Simple | Smart | Speed

CHOOLS

Regression Analysis

Learning Objectives

After reading this chapter, you should understand:

– What regression analysis is and what it can be used for.
– How to specify a regression analysis model.
– How to interpret basic regression analysis results.
– What the issues with, and assumptions of regression analysis are.
– How to validate regression analysis results.
– How to conduct regression analysis in SPSS.
– How to interpret regression analysis output produced by SPSS.

Keywords

Adjusted R2 • Autocorrelation • Durbin-Watson test • Errors • F-test • Heteroskedasticity • Linearity • Moderation • (Multi)collinearity • Ordinary least squares • Outliers • Regression analysis • Residuals • R2 • Sample size • Stepwise methods • Tolerance • Variance inflation factor • Weighted least squares

Agripro is a US-based firm in the business of selling seeds to farmers and distributors. Regression analysis can help them understand what drives customers to buy their products, helps explain their customer's satisfaction, and informs how Agripro measures up against their competitors. Regression analysis provides precise quantitative information on which managers can base their decisions.

7.1 Introduction

Regression analysis is one of the most frequently used tools in market research. In its simplest form, regression analysis allows market researchers to analyze relationships between one independent and one dependent variable. In marketing applications, the dependent variable is usually the outcome we care about (e.g., sales), while the independent variables are the instruments we have to achieve those outcomes with (e.g., pricing or advertising). Regression analysis can provide insights that few other techniques can. The key benefits of using regression analysis are that it can:

1. Indicate if independent variables have a significant relationship with a dependent variable.
2. Indicate the relative strength of different independent variables' effects on a dependent variable.
3. Make predictions.

Knowing about the effects of independent variables on dependent variables can help market researchers in many different ways. For example, it can help direct spending if we know promotional activities significantly increases sales.

Knowing about the relative strength of effects is useful for marketers because it may help answer questions such as whether sales depend more on price or on promotions. Regression analysis also allows us to compare the effects of variables measured on different scales such as the effect of price changes (e.g., measured in $) and the number of promotional activities.

Regression analysis can also help to make predictions. For example, if we have estimated a regression model using data on sales, prices, and promotional activities, the results from this regression analysis could provide a precise answer to what would happen to sales if prices were to increase by 5% and promotional activities were to increase by 10%. Such precise answers can help (marketing) managers make sound decisions. Furthermore, by providing various scenarios, such as calculating the sales effects of price increases of 5%, 10%, and 15%, managers can evaluate marketing plans and create marketing strategies.

7.2 Understanding Regression Analysis

In the previous paragraph, we briefly discussed what regression can do and why it is a useful market research tool. But what is regression analysis all about? To answer this question, consider Figure 7.1 which plots a dependent (y) variable (weekly sales in $) against an independent (x) variable (an index of promotional activities). Regression analysis is a way of fitting a "best" line through a series of observations. With "best" line we mean that it is fitted in such a way that it minimizes the sum of squared differences between the observations and the line itself. It is important to know that the best line fitted with regression analysis is not necessarily the true line (i.e., the line that holds in the population). Specifically, if we have data issues, or fail to meet the regression assumptions (discussed later), the estimated line may be biased.

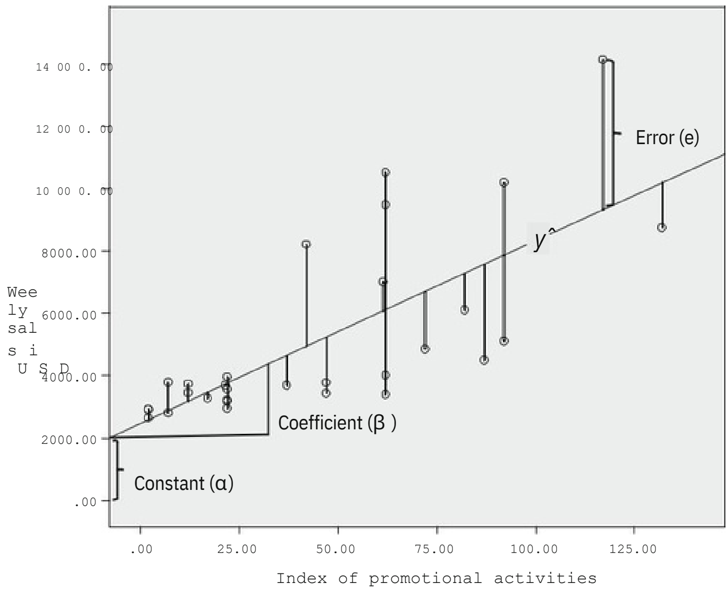


Fig. 7.1 A visual explanation of regression analysis

Before we introduce regression analysis further, we should discuss regression notation. Regression models are generally noted as follows:

$$y = \alpha + \beta_1 x_1 + e$$

What does this mean? The y represents the dependent variable, which is the variable you are trying to explain. In Fig. 7.1, we plot the dependent variable on the vertical axis. The α represents the constant (sometimes called intercept) of the regression model, and indicates what your dependent variable would be if all of the independent variables were zero. In Fig. 7.1, you can see the constant indicated on the y-axis. If the index of promotional activities is zero, we expect sales of around $2,500. It may of course not always be realistic to assume that independent variables are zero (just think of prices, these are rarely zero) but the constant should always be included to make sure that the regression model has the best possible fit with the data.

The independent variable is indicated by x_1. β_1 (pronounced as beta) indicates the (regression) coefficient of the independent variable x. This coefficient represents the gradient of the line and is also referred to as the slope and is shown in Fig. 7.1. A positive β_1 coefficient indicates an upward sloping regression line while a negative β_1 indicates a downward sloping line. In our example, the gradient slopes upward. This makes sense since sales tend to increase as promotional activities increase. In our example, we estimate β_1 as 55.968, meaning that if we increase promotional activities by one unit, sales will go up by $55.968 on average. In regression analysis, we can calculate whether this value (the β_1 parameter) differs significantly from zero by using a t-test.

The last element of the notation, the e denotes the error (or residual) of the equation. The term error is commonly used in research, while SPSS uses the term residuals. If we use the word error, we discuss errors in a general sense. If we use residuals, we refer to specific output created by SPSS. The error is the distance between each observation and the best fitting line. To clarify what a regression error is, consider Fig. 7.1 again. The error is the difference between the regression line (which represents our regression prediction) and the actual observation. The predictions made by the "best" regression line are indicated by $\hat{y}$ (pronounced y-hat). Thus, the error for the first observation is:[1]

$$e_1 = y_1 - \hat{y}_1$$

In the example above, we have only one independent variable. We call this bivariate regression. If we include multiple independent variables, we call this multiple regression. The notation for multiple regression is similar to that of bivariate regression. If we were to have three independent variables, say index of promotional activities (x_1), price of competitor 1 (x_2), and the price of competitor 2 (x_3), our notation would be:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

We need one regression coefficient for each independent variable (i.e., β_1, β_2, and β_3). Technically the βs indicate how a change in an independent variable influences the dependent variable if all other independent variables are held constant.[2]

Now that we have introduced some basics of regression analysis, it is time to discuss how to execute a regression analysis. We outline the key steps in Fig. 7.2. We first introduce the data requirements for regression analysis that determine if regression analysis can be used. After this first step, we specify and estimate the regression model. Next, we discuss the basics, such as which independent variables to select. Thereafter, we discuss the assumptions of regression analysis, followed by how to interpret and validate the regression results. The last step is to use the regression model, for example to make predictions.

7.3 Conducting a Regression Analysis

7.3.1 Consider Data Requirements for Regression Analysis

Several data requirements have to be considered before we undertake a regression analysis. These include the following:

- Sample size,
- Variables need to vary,
- Scale type of the dependent variable, and
- Collinearity.

[1] Strictly speaking, the difference between predicted and observed y-values is $\hat{e}$.

[2] This only applies to the standardized βs.

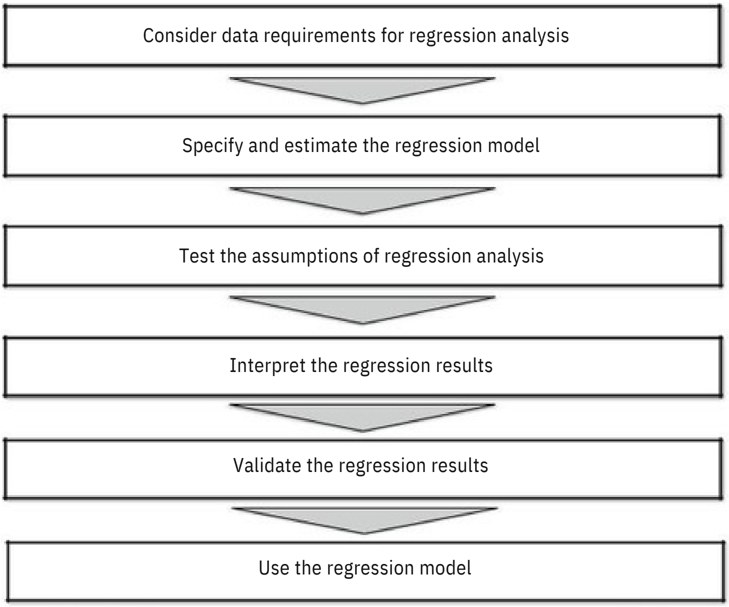


Fig. 7.2 Steps to conduct a regression analysis

7.3.1.1 Sample Size

The first data requirement is that we need a sufficiently large sample size. Acceptable sample sizes relate to a minimum sample size where you have a good chance of finding significant results if they are actually present, and not finding significant results if these are not present. There are two ways to calculate "acceptable" sample sizes.

– The first, formal, approach is a power analysis. As mentioned in Chap. 6 (Box 6.2), these calculations are difficult and require you to specify several parameters, such as the expected effect size or the maximum type I error you want to allow for to calculate the resulting level of power. By convention, 0.80 is an acceptable level of power. Kelley and Maxwell (2003) discuss sample size requirements.

– The second approach is through rules of thumb. These rules are not specific to a situation but are easy to apply. Green (1991) proposes a rule of thumb for sample sizes in regression analysis. Specifically, he proposes that if you want to test for individual parameters' effect (i.e., if one coefficient is significant or not), you need a sample size of 104 kþ. Thus, if you have ten independent variables, you need 104 þ 10 ¼ 114 observations.[3]

[3] Rules of thumb are almost never without issues. For Green's formula, these are that you need a larger sample size than he proposes if you expect small effects (an expected R2 of 0.10 or smaller). In addition, if the variables are poorly measured, or if you want to use a stepwise method, you need a larger sample size. With "larger" we mean around three times the required sample size if the expected R2 is low, and about twice the required sample size in case of measurement errors or if stepwise methods are used.

7.3.1.2 Variables Need to Vary

A regression model cannot be estimated if the variables have no variation. Specifically, if there is no variation in the dependent variable (i.e., it is constant), we also do not need regression, as we already know what the dependent variable's value is. Likewise, if an independent variable has no variation, it cannot explain any varia- tion in the dependent variable.

> No variation can lead to epic fails! Consider the admission tests set by the University of Liberia. Not a single student passed the entry exams. Clearly in such situations, a regression analysis will make no difference!
> http:// www. independent.co.uk/st udent/news /epic-fail-all-25000-student s-fail-university-entrance-exam-in-liberia-8785707.html

7.3.1.3 Scale Type of the Dependent Variable

The third data requirement is that the dependent variable needs to be interval or ratio scaled (scaling is discussed in Chap. 2). If the data are not interval or ratio scaled, alternative types of regression need to be used. You should use binary logistic regression if the dependent variable is binary and only takes on two values (e.g., zero and one). If the dependent variable consists of a nominal variable with more than two levels, you should use multinomial logistic regression. This should, for example, be used if you want to explain why people prefer product A over B or C. We do not discuss these different methods in this chapter, but they are intuitively similar to regression. For an introductory discussion of regression methods with dependent variables measured on a nominal scale, see Field (2013).

7.3.1.4 Collinearity

The last data requirement is that no or little collinearity is present. Collinearity is a data issue that arises if two independent variables are highly correlated. Multicol- linearity occurs if more than two independent variables are highly correlated. Perfect (multi)collinearity occurs if we enter two (or more) independent variables with exactly the same information in them (i.e., they are perfectly correlated).

> Perfect collinearity may happen because you entered the same independent variable twice, or because one variable is a linear combination of another (e.g., one variable is a multiple of another variable such as sales in units and sales in thousands of units). If this occurs, regression analysis cannot estimate one of the two coefficients and SPSS will automatically drop one of the independent variables.

In practice, however, weaker forms of collinearity are common. For example, if we study how much customers are wiling to pay in a restaurant, satisfaction with the waiter/waitress and satisfaction with the speed of service may be highly related. If this is so, there is little uniqueness in each variable, since both provide much the same information. The problem with having substantial collinearity is that it tends to disguise significant parameters as insignificant.

Fortunately, collinearity is relatively easy to detect by calculating the tolerance or VIF (Variance Inflation Factor). A tolerance of below 0.10 indicates that (multi) collinearity is a problem.[4] The VIF is just the reciprocal value of the tolerance. Thus, VIF values above ten indicate collinearity issues. We can produce these statistics in SPSS by clicking on Collinearity diagnostics under the Options button found in the main regression dialog box of SPSS.

You can remedy collinearity in several ways. If perfect collinearity occurs, SPSS will automatically delete one of the perfectly overlapping variables. SPSS indicates this through an additional table in the output with the title "Excluded Variables". If weaker forms of collinearity occur, it is up to you to decide what to do.

– The first option is to use factor analysis (see Chap. 8). Using factor analysis, you create a small number of factors that have most of the original variables' information in them but which are mutually uncorrelated. For example, through factor analysis you may find that satisfaction with the waiter/waitress and satisfaction with the speed of service fall under a factor called service satisfaction. If you use factors, collinearity between the original variables is no longer an issue.

– The second option is to re-specify the regression model by removing highly correlated variables. Which variables should you remove? If you create a correlation matrix (see Chap. 5) of all the independent variables entered in the regression model, you should focus first on the variables that are most strongly correlated. Initially, try removing one of the two most strongly correlated variables. Which one you should remove is a matter of taste and depends on your analysis set-up.

7.3.2 Specify and Estimate the Regression Model

To conduct a regression analysis, we need to select the variables we want to include and decide on how the model is estimated. In the following, we will discuss each step in detail.

7.3.2.1 Model Specification

Let's first show the main regression dialog box in SPSS to provide some idea of what we need to specify for a basic regression analysis. First open the dataset called

[4]The tolerance is calculated using a completely separate regression analysis. In this regression analysis, the variable for which the tolerance is calculated is taken as a dependent variable and all other independent variables are entered as independents. The R2 that results from this model is deducted from 1, thus indicating how much is not explained by the regression model. If very little is not explained by the other variables, (multi) collinearity is a problem.

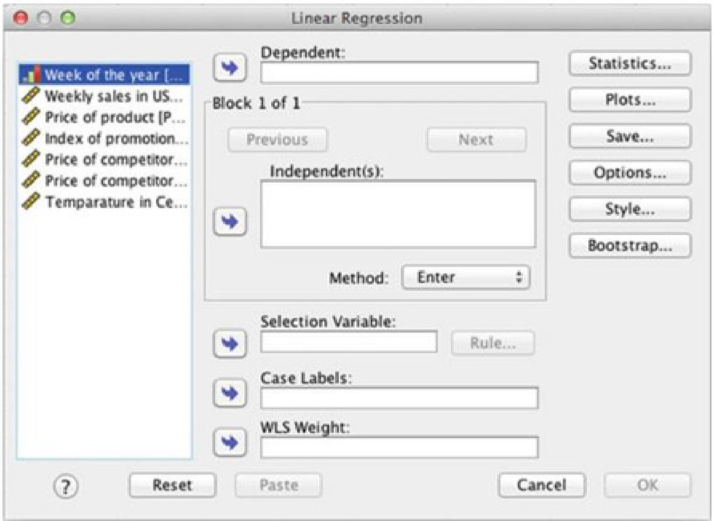


Fig. 7.3 The main regression dialog box in SPSS

Sales data.sav (8 Web Appendix ! Chap. 7). These data contain information on supermarket sales per week in $ (sales), the (average) price level (price), and an index of promotional activities (promotion), amongst other variables. After opening the dataset, click on u Analyze u Regression u Linear. This opens a box similar to Fig. 7.3.

For a basic regression model, we need to specify the Dependent variable and choose the Independent(s). As discussed before, the dependent variable is the variable we care about as the outcome.

How do we select independent variables? Market researchers usually select independent variables on the basis of what the client wants to know and on prior research findings. For example, typical independent variables explaining the super-market sales of a particular product include the price, promotional activities, level of in-store advertising, the availability of special price promotions, packaging type, and variables indicating the store and week. Market researchers may, of course, select different independent variables for other applications. A few practical suggestions to help you select variables:

– Never enter all the available variables at the same time. Carefully consider which independent variables may be relevant. Irrelevant independent variables may be significant due to chance (remember the discussion on hypothesis testing in Chap. 6) or can reduce the likelihood of determining relevant variables' significance.

– If you have a large number of variables that overlap in terms of how they are defined, such as satisfaction with the waiter/waitress and satisfaction with the speed of service, try to pick the variable that is most distinct or relevant to the client. Alternatively, you could conduct a factor analysis first and use the factor scores as input for the regression analysis (factor analysis is discussed in Chap. 8).

– Take the sample size rules of thumb into account. If practical issues limit the sample size to below the threshold recommended by the rules of thumb, use as few independent variables as possible. With larger sample sizes, you have more freedom to add independent variables, although they still need to be relevant.

As an example, we use sales as the dependent variable and price as well as the index of promotional activities as independent variables.

7.3.2.2 Model Estimation

Once we know which variables we want to include, we need to specify if all of them should be used, or if – based on the significance of the findings – the analysis procedure can further select variables from this set. There are two general options to select variables under Method in Fig. 7.3. Either you choose the independent variables to be in the model yourself (the enter method) or you let a process select the best subset of variables available to you (a stepwise method). There are many different types of stepwise methods such as the forward and backward methods, which we explain in Box 7.1.

Choosing between the enter and stepwise methods means making a choice between letting the researcher or a procedure choose the best independent variables. We recommend using the enter method. Why? Because stepwise methods often result in adding variables that are only significant "by chance," rather than truly interesting or useful. Another problem with forward and backward methods is related to how regression deals with missing values. Regression can only estimate models when it has complete information on all the variables. If a substantial number of missing values are present, using backward or forward methods may result in adding variables that are only relevant for a subset of the data for which

Box 7.1 Forward or backward methods

Forward and backward methods are often used for data mining purposes. How do these work? Starting with the constant (α) only, the forward method runs a very large number of separate regression models. Then it tries to find the best model by adding just one independent variable from the remaining variables. Subsequently it compares the results between these two models. If adding an independent variable produces a significantly better model, it proceeds by adding a second variable from the variables that remain. The resulting model (which includes the two independent variables) is then compared to the previous model (which includes one independent variable). This process is repeated until adding another variable does not improve the model significantly. The backward method does something similar but initially enters all variables that it may use and removes the least contributing independent variable until removing another makes the model significantly worse.

complete information is present. If data are missing, backward or forward methods often result in finding highly significant models that only use a small number of observations from the total number of available observations. In this case, the regression model fits a small set of the data well but not the entire data or population. Finally, as a market researcher, you want to select variables that are meaningful for the decision-making process. You also need to think about the actual research problem, rather than choosing the variables that produce the "best model." Does this mean that the forward and backward methods are completely useless? Certainly not! Market researchers commonly use stepwise methods to find their way around the data quickly and to develop a feel for relationships in the data.

After deciding on the variable selection process, we need to choose an estimation procedure. Estimation refers to how the "best line" we discussed earlier is calcu-lated. SPSS estimates regression models by default, using ordinary least squares (OLS). As indicated before, OLS estimates a regression line so that it minimizes the squared differences between each observation and the regression line. By squaring distances, OLS avoids negative and positive deviations from the regression line cancelling each other out. Moreover, by squaring the distances, OLS also puts greater weight on observations that are far away from the regression line. The sum of all these squared distances is called the sum of squares and is indicated in

A practical issue related to specifying and estimating the regression model is if we conduct just one regression analysis, or if we run multiple models. Market researchers typically run many different models. A standard approach is to start with relatively simple models, such as with one, two, or three independent variables. These independent variables should be those variables you believe are the most important ones. You should start with just a few variables because adding further variables may cause the already entered variables to lose significance. If important decisions are made with a regres-sion model, we need to be aware that sometimes variables that are significant in one model may no longer be significant if we add (or remove) variables. As discussed earlier, this is often due to collinearity. Once you have determined a number of key basic variables, you could (depending on the research purpose) add further independent variables until you have a model that satisfies your needs and does a good job of explaining the dependent variable. Generally, regression models have between 3 and 10 independent variables but bivariate regression models are also common. In specific applications, such as regres-sion models that try to explain economic growth, regression models can have dozens of independent variables.

the SPSS output. While minimizing the sum of squares, OLS also ensures that the mean of all errors is always zero. Because the error is zero on average, researchers sometimes omit the e from the regression notation. Nevertheless, errors do occur in respect of individual observations (but not on average). Figure 7.1

illustrates this. Almost all observations fall above or below the regression line. However, if we calculate the mean of all the squared distances of regression points above and below the line, the result is exactly zero. In Box 7.2 we discuss estima- tion methods other than OLS.

7.3.3 Test the Assumptions of Regression Analysis

We have already discussed several issues that determine if it is useful to run a regression analysis. We now turn to discussing the assumptions of regression analysis. If a regression analysis fails to meet the assumptions, regression analysis can provide invalid results. Four regression analysis assumptions are required to provide valid results:

Box 7.2 Different problems, different estimators

OLS is a very robust estimator. However, there are alternatives that work better and are best used in specific situations. Typically these situations occur if we violate on of the regression assumptions. For example, if the regression residuals are heteroskedastic, we need to use alternative procedures such as weighted least squares (WLS). We briefly discuss when WLS should be used in this chapter. If the expected mean error of the regression model is not zero, estimators such as two-staged least squares (2SLS) can be used in specific situations. If the errors are not independent, estimators such as random-effects estimators may be used. Such estimators are beyond the scope of this book. Greene's (2007) work discusses these, and other estimation procedures in detail.

1. The regression model can be expressed in a linear way,
2. The expected mean error of the regression model is zero,
3. The variance of the errors is constant (homoskedasticity), and
4. The errors are independent (no autocorrelation).

The fifth assumption is optional. If we meet this assumption, we have informa-

tion on how the regression parameters are distributed, thus allowing straightforward conclusions on their significance. If we fail to meet this assumption, the regression model will still be accurate but it becomes more difficult to determine the regression parameters' significance.

5. The errors need to be approximately normally distributed.

We next discuss these assumptions and how we can test each of them.

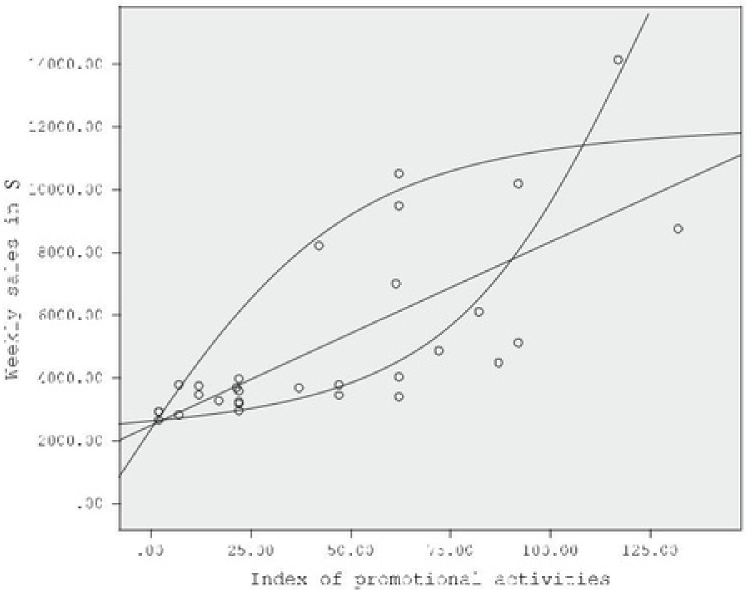


Fig. 7.4 Different relationships between promotional activities and weekly sales

7.3.3.1 First Assumption: Linearity

The first assumption means that we can write the regression model as $y = \alpha + \beta_1 x_1 + e$. Thus, relationships such as $\beta_1^2 x$ are not permissible. On the other hand, expressions such as x_1^2 or $\log(x_1)$ are possible as the regression model is still specified in a linear way. As long as you can write a model where the regression parameters (the βs) are linear, you satisfy this assumption. A separate issue is if the relationship between an independent variable x and the dependent variable y, is linear. Checking the linearity between x and y variables can be done by plotting the independent variables against the dependent variable. Using a scatter plot, we can then assess whether there is some type of non-linear pattern. Figure 7.4 shows such a plot. The straight line indicates a linear relationship. For illustration purposes, we have also added an upward sloping and downward sloping line. The upward sloping line corresponds to an x_1^2 transformation, while the downward sloping line corresponds to a $\log(x_1)$ transformation. For this particular data, it appears however that a linear line fits the data best. It is important to correctly specify the relation-ship, because if we specify a relationship as linear when it is in fact non-linear, the regression analysis results do not fit the data in the best possible way. After transforming x_1 by squaring it or taking the log, you still satisfy the assumption of specifying the regression model in a linear way, despite that the relationship between x and y is nonlinear.

7.3.3.2 Second Assumption: Expected Mean Error is Zero

The second assumption is that the expected (not the estimated!) mean error is zero. If we do not expect the sum of the errors to be zero, we obtain a line that is biased. That is, we have a line that consistently over- or under-estimates the true relationship. This assumption is not testable by means of statistics, as OLS always renders a best line where the mean error is exactly zero. If this assumption is challenged, this is done on theoretical grounds. Imagine that we want to explain the weekly sales in $ of all US supermarkets. If we were to collect our data only in downtown areas, we would mostly sample smaller supermarkets. Thus, a regression model fitted using the available data would differ from those models obtained if we were to include all supermarkets. Our error in the regression model (estimated as zero) therefore differs from the population as a whole (where the estimate should be truly zero). Further-more, omitting important independent variables could cause the expected mean not to be zero. Simply put, if we

were to omit a relevant variable x2 from a regression model that only includes x1, we induce a bias in the regression model. More precisely, β1 is likely to be inflated, which means that the estimated value is higher than it should actually be. Thus, β1 itself is biased because we omit x2!

7.3.3.3 Third Assumption: Homoskedasticity

The third assumption is that the errors' variance is constant, a situation we call homoskedasticity. Imagine that we want to explain the weekly sales of various supermarkets in $. Clearly, large stores have a much larger spread in sales than small supermarkets. For example, if you have average weekly sales of $50,000, you might see a sudden jump to $60,000 or a fall to $40,000. However, a very large supermarket could see sales move from an average of $5,000,000–$7,000,000. This issue causes weekly sales' error variance to be much larger for large supermarkets than for small supermarkets. We call this non-constant variance heteroskedasticity. We visualize the increasing error variance of supermarket sales in Fig. 7.5, in which we can see that the errors increase as weekly sales increase.

If we estimate regression models on data in which the variance is not constant, they will still result in errors that are zero on average (i.e., our predictions are still correct), but this may cause some βs not to be significant, whereas, in reality, they are.

Unfortunately, there is no easy (menu-driven) way to test for heteroskedasticity in SPSS. Thus, understanding whether heteroskedasticity is present, is (if you use the SPSS menu functions) only possible on theoretical grounds and by creating graphs. On theoretical grounds, try to understand whether it is likely that the errors increase as the value of the dependent variable increases or decreases. If you want to visualize heteroskedasticity, it is best to plot the errors against the dependent variable, as in Fig. 7.5. As the dependent variable increases or decreases, the variance should appear as constant. If heteroskedasticity is an issue, the points are often funnel shaped, becoming more, or less, spread out across the graph. This funnel shape is typical of heteroskedasticity and indicates increasing variance across the errors.

If you think heteroskedasticity is an issue, SPSS can deal with it by using weighted least squares (WLS). Simply use the variable that you think causes the error variance not to be constant (e.g., store size) and "weight" the results by this variable. In Fig. 7.3 you see a box labelled WLS Weight at the bottom to which you can add the variable that causes the increase in error variance. Only use WLS if

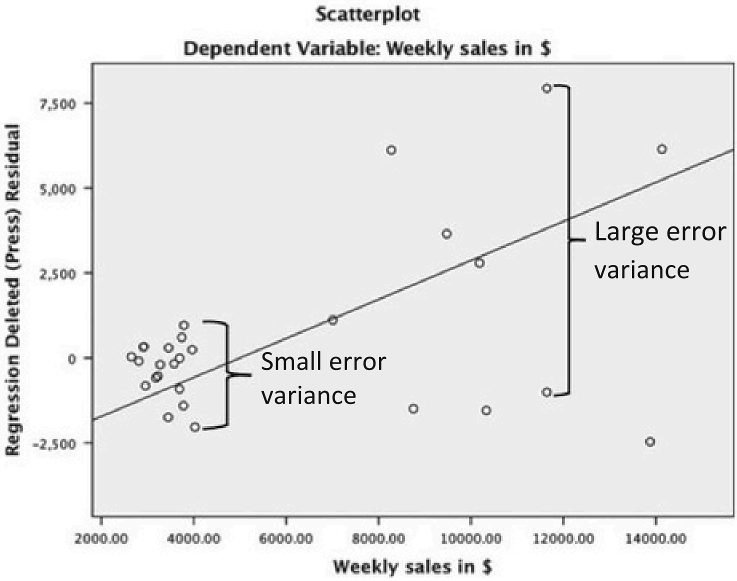


Fig. 7.5 An example of heteroskedasticity

heteroskedasticity is a real concern. If WLS is used, and heteroskedasticity is no problem, or the weight variable has not been chosen correctly, the regression results may be invalid.

7.3.3.4 Fourth Assumption: No Autocorrelation

The fourth assumption is that the regression model errors are independent; that is, the error terms are uncorrelated for any two observations. Imagine that you want to explain the sales of a particular supermarket using that supermarket's previous week sales. It is very likely that if sales increased last week, they will also increase this week. This may be due to, for example, a growing economy, or other reasons that underlie supermarket sales growth. This issue is called autocorrelation and means that regression errors are correlated positively, or negatively, over time. Fortunately, we can identify this issue using the Durbin–Watson test. The Durbin– Watson test assesses whether there is autocorrelation by testing a null hypothesis of no autocorrelation, which is tested against a lower and upper bound for negative autocorrelation and against a lower and upper bound for postive autocorrelation. Thus there are four critical values. If we reject the null hypothesis of no autocorre-lation, we find support for an alternative hypothesis that there is some degree of autocorrelation. To carry out this test, first sort the data on the variable that indicates the time dimension in your data, if you have this included in your data. Otherwise, the test should not be carried out. With time dimension, we mean that you have at least two observations collected from a single respondent or object at different points in time. Do this by going to u Data u Sort Cases. Then enter your time variable under Sort by: and click on OK. To carry out the actual test, you need to check Durbin–Watson under the Statistics option of the main regression dialog box in SPSS. SPSS calculates a Durbin–Watson statistic, but does not indicate if the test is significant or not. This requires comparing the calculated Durbin–Watson value with the critical Durbin–Watson value. These Durbin–Watson values lie between 0 and 4. Essentially, there are four situations. First, the errors may be positively related (called positive autocorrelation). This means that if we take observations ordered according to time, we observe that positive errors are typically followed by positive errors and that negative errors are typically followed by negative errors. For example, supermarket sales usually increase over certain periods in time (e.g., before Christmas) and decrease in other periods (e.g., the summer holidays). Second, if positive errors are commonly followed by negative errors and vice-versa, we have negative autocorrelation. Negative autocorrelation is less common than positive autocorrelation, but also occurs. If we study, for example, how much time salespeople spend on shoppers, we may see that if they spend much time on one shopper, they spend less time on the next, allowing the salesperson to stick to his/her schedule or simply go home on time. Third, if no systematic pattern of errors occurs, we have no autocorrelation. Fourthly, the Durbin-Watson values may fall in between the lower and upper critical value. In this case, the test is inconclusive. We indicate these four situations in Fig. 7.6. Which situation occurs, depends on the interplay between the

Durbin-Watson test statistic (d) and the lower (dL) and upper (dU) critical value.

– If the test statistic is lower than the lower critical value (d < dL) we have positive autocorrelation.
– If the test statistic is higher than 4 minus the lower critical value (d > 4-dL) we have negative autocorrelation
– If the test statistic falls between the upper critical value and 4 minus the upper critical value (dU < d < 4-dU) we have no autocorrelation.
– If the test statistic falls in-between the lower and upper critical value (dl < d < dU) or it falls in-between 4 minus the upper critical value and 4 minus the lower critical value (4-dU < d < 4-dL) we cannot make a decision on the presence of autocorrelation.

The critical values can be found on the website accompanying this book (8 Web Appendix ! Chap. 7). From this table, you can see that the lower critical value of a model with five independent variables and 200 observations is 1.718 and the upper critical value is 1.820. Figure 7.6 shows the intervals for the above example and if the Durbin–Watson test concludes that there is no autocorrelation, you can proceed with the regression model. If the Durbin–Watson test indicates autocorrelation, you may have to use models that account for this problem, such as panel and time-series

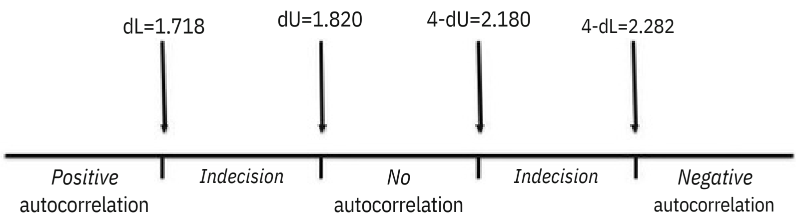


Fig. 7.6 Durbin-Watson test values (n ¼ 200, k ¼ 5)

models. We do not discuss these in this book, but a useful source of further information is Hill et al. (2008).

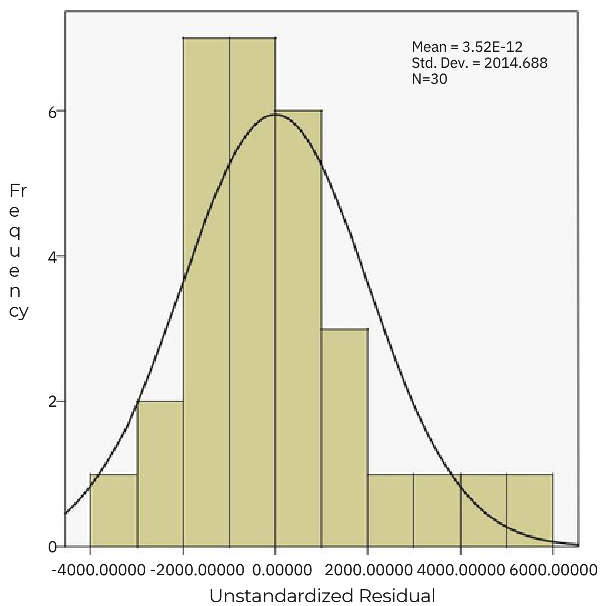
7.3.3.5 Fifth (Optional) Assumption: Error Distribution

The fifth, optional, assumption is that the regression model errors are approximately normally distributed. If this is not the case, the t-values may be incorrect. However, even if the errors of the regression model are not normally distributed, the regres-sion model still provides good estimates of the coefficients. Therefore, we consider this assumption as optional. Potential reasons for non- normality include outliers (discussed in Chap. 5) and a non-linear relationship between an independent and a dependent variable.

There are two main ways of checking for normally distributed errors, either you use plots or you can perform a formal test. The plots are easily explained and interpreted and may suggest the source of non-normality (if present). The formal test may indicate non-normality and provide absolute standards. However, the formal test results reveal little about the source of non-normality.

To test for non-normality using plots, first save the unstandardized errors by going to the Save dialog box in the regression menu. Then, create a histogram of these errors and plot a normal curve in it to understand if any deviations from normality are present. We can make histograms by going to u Graphs u Legacy Dialogs u Histogram. Make sure to check Display normal curve. The result may look something like Fig. 7.7. How do we interpret this figure? If we want the errors to be approximately normally distributed, the bars should end very “close” to the normal curve, which is the black bell-shaped curve. What “close” means exactly is

Fig. 7.7 Histogram of the errors



open to different interpretations, but Fig. 7.7 suggests that the errors produced by the estimated regression model are almost normally distributed.

In addition to the normal curve, SPSS produces a table, showing the results of two formal tests of normality (i.e., Kolmogorov-Smirnov and Shapiro-Wilk). Since we have only 30 observations in our dataset, we should use the Shapiro–Wilk test (see Chap. 6) as a formal test of normality. As we can easily see, the Shapiro–Wilk test result indicates that the errors are normally distributed as we cannot reject the null hypothesis at a significance level of 5% (p-value ¼ 0.084) (Table 7.1).

7.3.4 Interpret the Regression Results

In the previous sections, we discussed how to specify a basic regression model and how to test regression assumptions. We now turn to discussing the fit of the regression model, followed by the interpretation of the effects of individual variables.

Table 7.1 Output produced by the Shapiro–Wilk test

Tests of Normality

	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
UnstandardizedResidual	.131	30	.200*	.939	30	.084

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

7.3.4.1 Overall Model Fit

We can assess the overall model fit using the (adjusted) R2 and significance of the F-value.

The R2 (or coefficient of determination) indicates the degree to which the model explains the observed variation in the dependent variable, relative to the mean. In Fig. 7.8, we explain this graphically with a scatter plot. The y-axis relates to the dependent variable (weekly sales in $) and the x-axis to the independent variable (price). In the scatter plot, we see 30 observations of sales and price (note that we use a small sample size for illustration purposes). The horizontal line (at about $5,000 sales per week) refers to the average sales of all 30 observations. This is also our benchmark. After all, if we were to have no regression line, our best estimate of the weekly sales is also the average. The sum of all squared differences between each observation and the average is the total variation or total sum of the squares (usually referred to as SST). We indicate the total variation for only one observation on the right of the scatter plot.

The upward sloping line (starting at the y-axis at about $2,500 sales per week when there are no promotional activities) is the regression line that is estimated

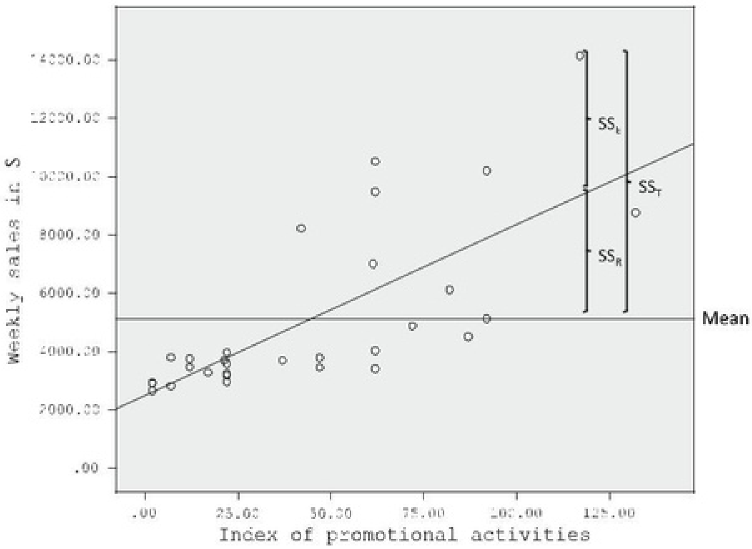


Fig. 7.8 Explanation of the R^2

using OLS. If we want to understand what the regression model adds beyond the average (the benchmark), we can calculate the difference between the regression line and the average. We call this the regression sum of the squares (usually abbreviated SSR) as it is the variation in the data that is explained by the regression analysis. The final point we need to understand regarding how well a regression line fits the available data, is the unexplained sum of squares. This refers to the regression error that we discussed previously and which is consequently denoted as SSE. In more formal terms, we can describe these types of variation as follows:

$$SST = SSR + SSE$$

This is the same as:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Here, n describes the number of observations, y_i is the value of the independent variable for observation i, $\hat{y}_i$ is the predicted value of observation i and y is the mean value of y. As you can see, this description is very similar to the one-way ANOVA, discussed in Chap. 6. A good regression line should explain a substantial amount of variation (have a high SSR) relative to the total variation (SST). This is the R^2 and we can calculate this as:

$$R^2 = \frac{SSR}{SST}$$

The R2 always lies between 0 and 1, where a higher R2 indicates a better model fit. When interpreting the R2, higher values indicate that more of the variation in y is explained by variation in x, and therefore that the SSE is low relative to the SSR.

> It is difficult to provide rules of thumb regarding what R2 is appropriate, as this varies from research area to research area. For example, in longitudinal studies R2s of 0.90 and higher are common. In cross-sectional designs, values of around 0.30 are common while for exploratory research, using cross-sectional data, values of 0.10 are typical. In scholarly research that focuses on marketing issues, R2 values of 0.75, 0.50, or 0.25 can, as a rough rule of thumb, be respectively described as substantial, moderate, or weak.

If we use the R2 to compare different regression models (but with the same dependent variable), we run into problems. If we add irrelevant variables that are slightly correlated with the dependent variable, the R2 will increase. Thus, if we use the R2 as the only basis for understanding regression model fit, we are biased towards selecting regression models with many independent variables. Selecting a model only based on the R2 is plainly not a good strategy, as we want regression models that do a good job of explaining the data (thus a low SSE), but which also have few independent variables (these are called parsimonious models). We do not want too many independent variables because this makes using the regression model more difficult. It is easier to recommend that management changes a few key variables to improve an outcome than to recommend a long list of somewhat related variables. Of course, relevant variables should always be included. To quote Albert Einstein: “Everything should be made as simple as possible, but not simpler!”

To avoid a bias towards complex models, we can use the adjusted R2 to select regression models. The adjusted R2 only increases if the addition of another independent variable explains a substantial amount of variance. We calculate the adjusted R2 as follows:

$$\text{Adjusted}R^2 = 1 - (1 - R^2) \cdot \frac{(n-1)}{n-k-1}$$

Here, n describes the number of observations and k the number of independent variables (not counting the constant α). This adjusted R2 is a relative measure and should be used to compare different regression models with the same dependent variable. You should pick the model with the highest adjusted R2 when comparing regression models.

However, do not blindly use the adjusted R2 as a guide, but also look at each individual variable and see if it is relevant (practically) for the problem you are researching. Furthermore, it is important to note that we cannot interpret the adjusted R^2 as the percentage of explained variance in the sample used for regression analysis. The adjusted R^2 is only a measure of how much the model explains while controlling for model complexity.

Besides the (adjusted) R^2, the F-test is an important determinant of model fit. The test statistic's F-value is the result of a one-way ANOVA (see Chap. 6) that tests the null hypothesis that all regression coefficients together are equal to zero. Thus, the following null hypothesis is tested:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = 0$$

The alternative hypothesis is that at least one β differs from zero. If the regression coefficients were all equal to zero, then the effect of all the independent variables on the dependent variable is zero. In other words, there is no (zero) relationship between the dependent variable and the independent variables. If we do not reject the null hypothesis, we need to change the regression model or, if this is not possible, report that the regression model is insignificant.

The test statistic's F-value closely resembles the F-statistic, as discussed in Chap. 6 and is also directly related to the R^2 we discussed previously. We can calculate the F-value as follows:

$$F = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{R^2}{1-R^2} \cdot \frac{(n-k-1)}{k}$$

The test statistic follows an F-distribution with k and (n – k – 1) degrees of freedom. Finding that the p-value of the F-test is below 0.05 (i.e., a significant model), does not, however, automatically mean that all of our regression coefficients are significant or even that one of them is significant, when considered in isolation. However, if the F-value is significant, it is highly likely that at least one or more regression coefficients are significant.

When we interpret the model fit, the F-test is the most critical, as it determines if the overall model is significant. If the model is insignificant, we do not interpret the model further. If the model is significant, we proceed by interpreting individual variables.

7.3.4.2 Effects of Individual Variables

After having established that the overall model is significant and that the R^2 is satisfactory, we need to interpret the effects of the various independent variables used to explain the dependent variable. First, we need to look at the t-values reported for each individual parameter. These t-values are similar to those discussed in Chap. 6 in the context of hypothesis testing. If a regression coefficient's p-value (indicated in SPSS by the column headed by Sig.) is below 0.05, we generally say that that particular independent variable relates significantly to the dependent variable.

To be precise, the null and alternative hypotheses tested for an individual parameter (e.g., β_1) are:

H0:β1¼0

H1:β1¼60

If a coefficient is significant (meaning we reject the null hypothesis), we also need to look at the unstandardized and standardized β coefficients. The unstandard-ized β coefficient indicates the effect of a 1-unit increase in the independent variable (on the scale in which the original independent variable is measured) on the dependent variable. Thus it is the partial relationship between a single independent variable and the dependent variable. At the very beginning of this chapter, we learned that there is a positive relationship between promotional activities and the weekly sales in $ with a

β1 coefficient of 55.968. This means that if we increase promotional activities by one unit, weekly sales are expected to go up by $55.968. In other variables, the effect could of course be negative (e.g., increasing prices reduces sales). Importantly, if we have

multiple independent variables, the unstan-dardized β1 coefficient is the effect of an increase of that independent variable by one unit, keeping the other independent variables constant.

While this is a very simple example, we might run a multiple regression in which the independent variables are measured on different scales, such as in $, units sold, or on Likert scales. Consequently, the independent variables' effects cannot be directly compared with one another as their influence also depends on the type of scale used. Comparing the (unstandardized) β coefficients would in any case amount to comparing apples with oranges!

Fortunately, the standardized βs allow us to compare the relative effect of differently measured independent variables. This is achieved by expressing β as standard deviations with a mean of zero. The standardized βs coefficient expresses the effect of a single standardized deviation change of the independent variable on the dependent variable. All we need to do is to look at the highest absolute value. This value indicates which variable has the strongest effect on the dependent variable. The second highest absolute value indicates the second strongest effect, etc. Only consider the significant βs in this respect, as insignificant βs do not (statistically) differ from zero! Practically, the standardized β is important, because it allows us to ask questions on what, for example, the relative effect of promotional activities is relative to decreasing prices. It can therefore guide management decisions.

While the standardized βs are helpful from a practical point of view, there are two issues. First, standardized βs allow comparing the coefficients only within and not between models! Even if you add just a single variable to your regression model, standardized βs may change substantially. Second, standardized βs are not meaningful when the independent variable is binary.

When interpreting (standardized) β coefficients, you should always keep the effect size in mind. If a β coefficient is significant, it indicates merely an effect that differs from zero. This does not necessarily mean that the effect is managerially relevant. For example, we may find a $0.01 sales effect of spending $1 more on promotional activities that is statistically significant. Statistically, we could conclude that the effect of a $1 increase in promotional activities increases sales by an average of $0.01 (just one dollar cent). While this effect differs significantly from zero, in practice we would probably not recommend increasing promotional activities (we would lose

money at the margin) as the effect size is just too small.[5]

There are also situations in which an effect is not constant for all observations but depends on the values of another variable. To disclose such effects, researchers can run a moderation analysis, which we discuss in Box 7.3.

Box 7.3 Moderation

The discussion on the effects of individual variables assumes that there is only one effect. That is, there is only one β parameter that represents all observations well. Often, this is not true. For example, the link between customer satisfaction and loyalty has been shown to be stronger for people with low income than for people with high income. In other words, there is heterogeneity in the effect between satisfaction and loyalty.

Moderation analysis is one way to test if such heterogeneity is present. A moderator variable, usually denoted with m, is a variable that changes the strength (or even direction) of the relationship between the independent variable (x) and the dependent variable (y). This moderation variable is frequently called an interaction variable. The moderating variable can weaken or strengthen the effect of x on y. Potentially, the m variable could even reverse the effect of x on y.

Moderation is easy to test if the moderator variable m is binary, ordinal, or interval scaled. All that is required is to create a new variable that is the multiplication of x m. This can be done in SPSS using u Transform u Compute. The regression model then takes the following form:

y¼αþβxþβmþβx mþe

In words, conducting a moderator analysis requires entering the independent variable x, the moderator variable m and the product x·m. After estimating this

(continued)

[5] An interesting perspective on significance and effect sizes is offered by Cohen's (1994) classical article "The Earth is Round (p <.05).

Box 7.3 (continued)

regression model, you can interpret the significance and sign of the β3 parameter. A significant effect suggests that:

– The effect of x increases as m increases (when the sign of β3 is positive),
– The effect of x decreases as m increases (when the sign of β3 is negative).

Finding a significant moderator effect suggests heterogeneity in the effect

of x on y, where the effect of x of y may increase as m increases or may decrease as m decreases.
For a further discussion on moderation analyses, please see David Kenny's discussion on moderation (http://www.davidakenny.net/cm/moderation.htm) or the advanced discussion of Aiken and West (1991). Jeremy Dawson's website (http://www.jeremydawson.co.uk/slopes.htm) offers a tool to visualize moderation effects. An example of a moderation analysis is for example found in (Mooi and Frambach 2009).

7.3.5 Validate the Regression Model

After we have checked for the assumptions of regression analysis and interpreted the results, we need to check for the stability of the regression model. Stability means that the results are stable over time, do not vary across different situations, and do not depend heavily on the model's specification. We can check for the stability of a regression model in several ways.

1. We could validate the regression results by splitting our data into two parts (called split-sample validation) and run the regression model again on each subset of data. 70% of the randomly chosen data are often used to estimate the regression model and the remaining 30% are used for comparison purposes. We can only split the data if the remaining 30% still meets the sample size rules of thumb discussed earlier. If the use of the two samples results in similar effects, we can conclude that the model is stable.

2. We can also cross-validate our findings on a new dataset and see if those findings are similar to the original findings. Again, similarity in the findings indicates stability and that our regression model is properly specified. This, naturally, assumes that we have a second dataset.

3. We could add a number of alternative variables to the model. But we would need to have more variables available than included in the regression model to do so. For example, if we try to explain weekly supermarket sales, we could use a number of "key" variables (e.g., the breadth of the assortment or downtown/non-downtown location) in our regression model to help us. Once we have a suitable regression model, we could use these variables. If the basic findings of, for example, promotional activities are the same for stores with a differing assort-ment width or store location (i.e., the assortment width and location are not

significant), we conclude that the effects are stable. However, it might also be the opposite, but whatever the case, we want to know.

Note that not all regression models need to be identical when you try to validate the results. The signs of the individual parameters should at least be consistent and significant variables should remain so, except if they are marginally significant, in which case changes are expected (e.g., $p = 0.045$ becomes $p = 0.051$).

7.3.6 Use the Regression Model

When we have found a useful regression model that satisfies the assumptions of regression analysis, it is time to use the regression model. A key use of regression models is prediction. Essentially, prediction entails calculating the values of the dependent variable based on assumed values of the independent variables and their related but previously calculated unstandardized β coefficients. Let us illustrate this by returning to our opening example. Imagine that we are trying to predict weekly supermarket sales (in $) (y) and have estimated a regression model with two independent variables: the average price (x_1) and an index of promotional activities (x_2). The regression model for this is as follows:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

If we estimate this model on a dataset, the estimated coefficients using regression analysis could be similar to those in Table 7.2.

Table 7.2 Table containing sample regression coefficients[a]

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.Error	Beta		
1	(Constant)	29011.585	18448.456		1.573	.127
	Priceofproduct	−24003.037	16694.676	−.241	−1.438	.162
	Indexofpromotional activities	44.227	13.567	.547	3.260	.003

a. Dependent Variable: Weekly sales in USD

We can also use these coefficients to make predictions of sales in different situations. Imagine, for example, that we have set the price at $1.10 and promo- tional activities at 50. Our expectation of the weekly sales would then be:

$$\hat{y} = 29{,}011.585 - 24{,}003.037 \quad \$1.10 + 44.227 \quad 50\ \text{promotional activities}$$
$$= \$4{,}819.594 \text{ sales per week.}$$

We could also build several scenarios to plan for different situations, by, for example, increasing the price to $1.20 and reducing promotional activities to 40. Regression models can be used like this to, for example, automate stocking and logistical planning or develop strategic marketing plans.

Another way in which regression can help is by providing insight into variables' specific effects. For example, if the effect of price is not significant, it may tell managers that the supermarket's sales are relatively insensitive to pricing decisions. Alternatively, the strength of promotional activities' effect may help managers understand whether promotional activities are useful.

Table 7.3 summarizes (on the left side) the major theoretical decisions we need to make if we want to run a regression model. On the right side, these decisions are then "translated" into SPSS actions, which are related to these theoretical decisions.

Table 7.3 Key steps involved in carrying out a regression analysis

Theory	ExecutioninSPSS
Issues with regression analysis	
Isthesamplesizesufficient?	Conductapoweranalysis.Alternatively,checkif sample size is 104 þ k, where k indicates the number of independent variables.
Dothedependentandindependent variablesshowvariation?	Calculatethestandarddeviationofthevariablesby goingtouAnalyzeuDescriptiveStatisticsu Descriptives u Options (check Std. Deviation). At the very least, the standard deviation should be a positive value. UseChap.2todeterminethemeasurementlevel.
Isthedependentvariableintervalor ratio scaled?	
Is(multi)collinearitypresent?	CheckfortoleranceandVIF.DothiswithuAnalyze u Regression u Linear u Statistics (check Collinearity diagnostics). The tolerance should be above 0.10. The VIF should be below 10.
Specifying and estimating the regression model	
Select variables based on theory or based on strength of effects	Preferably use the enter method. If stepwise methods are used (such as the forward method), only add variables that could have a relationship with the dependent variable.
Testing the assumptions of regression analysis	
Is the relationship between the independent and dependent variables linear?	Consider whether you can write the regression model asy¼αþβ1x1þ...þβkxkþe.
	To understand if the independent variables are linearly related to the dependent variable, plot the y variables separately against the dependent variable of

(continued)

Table 7.3 (continued)

Theory	Execution in SPSS the regression model. Create scatter
	plots using u Graphs u Legacy Dialogs u Scatter/Dot (choose Simple Scatter). If you see a non-linear pattern showing up, non-linearity is an issue. To specify a different relationship, see the transform variables section in Chap. 5.
Is the expected mean error of the regression model zero?	No actions in SPSS. Choice made on theoretical grounds.
Are the errors constant (homoskedastic)?	Plot the residual of the regression model on the y-axis and the dependent variable on the x-axis, using a scatter plot under u Graphs u Legacy Dialogs u Scatter/Dot (choose Simple Scatter). If you see that the errors in/decrease as the dependent variable increases, the variance of the errors is not constant. You can use WLS to remedy this.
Are the errors correlated (autocorrelation)?	First assess if there is a time component to the data (i.e., multiple observations, across time, from one respondent/object). If there is, sort the data according to the time variable and conduct the Durbin–Watson test. Compare the calculated Durbin–Watson test statistic with the critical lower and upper values. If positive or negative autocorrelation is present, panel or time-series models need to be used: u Analyze u Regression u Linear u Statistics and check the Durbin–Watson box.
Are the errors normally distributed?	Create a histogram of the errors with a standard normal curve in it: u Graphs u Legacy Dialogs u Histogram and enter the saved errors. Also check Display normal curve. Calculate the Kolmogorov-Smirnov test (for n 50) or Shapiro–Wilk test (for n < 50). u Analyze u Descriptive Statistics u Explore u Plots and check the Normality plots with tests box.
Interpret the regression model	
Consider the overall model fit.	Check the R2 and significance of the F-value. For model comparisons, use the adjusted R2.
Consider the effects of the independent variables separately.	Check the (standardized) β. Also check the sign of the β. Consider significance of the t-value.
Validate the model	
Are the results robust?	Split the file into subsets or run the regression model against another sample to check for robustness. Add additional variables that may be useful and check if a similar regression model results.

7.4 Example

In the example, we take a closer look at the American Customer Satisfaction Index (ACSI Data.sav, 8 Web Appendix ! Chap. 7). Every year, the American Cus- tomer Satisfaction Index (ACSI) surveys about 80,000 Americans about their level of satisfaction with a number of products and services. These satisfaction scores are used to benchmark competitors and to rate industries. For example, towards the beginning of 2014, the Quaker (PepsiCo), the H.J. Heinz Company, and General Mills were rated as the three food manufacturers with the highest scores. If you go to http://www.theacsi.org, you will find the current scores for various industries.[6] The ACSI data contain several variables, but we only focus on the following (variable names in parentheses):

– Overall Customer Satisfaction (lvsat) is measured by statements put to consumers about their overall satisfaction, expectancy disconfirmation (degree to which performance falls short of, or exceeds, expectations) and performance versus the customer's ideal product or service in the category.
– Customer Expectations (lvexpect) is measured by statements put to consumers about their overall expectations of quality (prior to purchase), their expectation regarding to how well the product fits the customer's personal requirements (prior to purchase), and expectation regarding reliability, or how often things will go wrong (prior to purchase).
– Perceived Value (lvvalue) is the consumers' rating of quality given price, and price given quality.
– Customer Complaints (lvcomp) captures whether or not the customer has complained formally or informally about the product or service (1 ¼ yes, 0 ¼ no).

The data includes 1,640 responses from customers, but due to item non-response, the actual number of responses for each variable is lower.

7.4.1 Consider Data Requirements for Regression Analysis

First we need to check if our sample size is sufficient. By calculating descriptive statistics (u Analyze u Descriptive Statistics u Descriptives; see Chap. 5) of the four above mentioned variables we can see that we have 1,640 valid listwise observations. This means that we have complete information for 1,640 observations. This is far above the minimum sample sizes as recommended by Green (1991). The sampling process has been documented by Fornell, Johnson, Anderson, Cha, and Bryant (1996) and we assume this is done correctly. Looking at the dependent and independent variables' variance, we can also see that all variables show variation. Finally, as our dependent variable is also ratio scaled, we can proceed with regression analysis. Multicollinearity might be an issue, but

[6] For an application of the ACSI, see, for example, Ringle et al. (2010).

we can only check this thoroughly after running a regression analysis. We will therefore discuss this aspect later.

7.4.2 Specify and Estimate the Regression Model in SPSS

Although it is useful to know who comes out on top, from a marketing perspective, it is more useful to know how organizations can increase their satisfaction. We can use regression for this and explain how a number of independent variables relate to satisfaction. Simply click on u Analyze u Regression u Linear and then enter Overall Customer Satisfaction into the Dependent box and Customer Expectations, Perceived Value, and Customer Complaints into the Independent(s) box. Figure 7.9 shows the regression dialog box in SPSS.

SPSS provides us with a number of options. Under Method choose Enter. The enter option includes all variables added into the Independent(s) box and does not

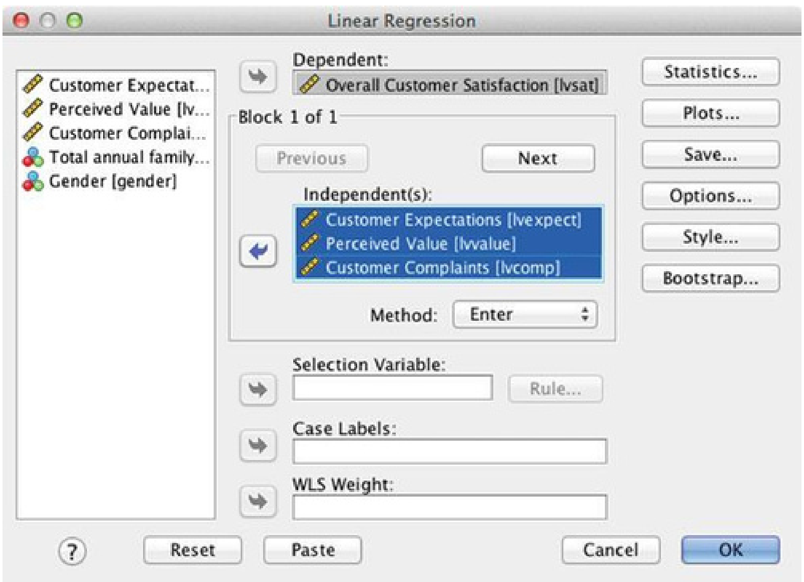


Fig. 7.9 The Linear Regression dialog box

remove any of the variables on statistical grounds (as opposed to the stepwise methods). Under Statistics in the main regression dialog box (see Fig. 7.10), SPSS offers several options on the output that you may want to see. The Estimates and Model fit options are checked by default and are essential. The Confidence intervals and Covariance matrix options are not necessary for standard analysis

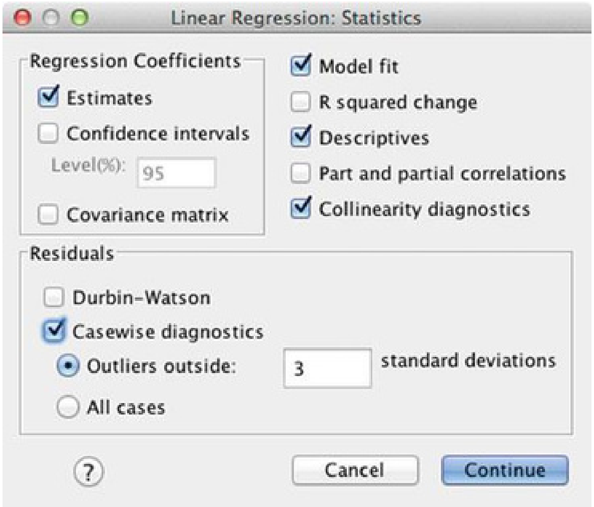


Fig. 7.10 The statistics dialog box

purposes, so we skip these. The R squared change option is only useful if you select any of the stepwise methods but is irrelevant if you use the (recommended) enter option. The Descriptives option does what it says and provides the mean, standard deviation, and number of observations for the dependent and independent variables. The Part and partial correlations option produces a correlation matrix, while the Collinearity diagnostics option checks for (multi)collinearity. The Durbin–Watson option checks for autocorrelation, while Casewise diagnostics provides outlier diagnostics. In this case, there is no time component to our data and thus the Durbin–Watson test is not applicable.

Next, make sure the Estimates, Model fit, Descriptives, Collinearity diagnostics, and Casewise diagnostic options are checked. Then click on Con-tinue. In the main regression dialog box, click on Save. This displays a dialog box similar to Fig. 7.11. Here, you can save predicted values and residuals.

Check the boxes Unstandardized under Predicted Values and Residuals. After clicking on Continue in the Linear Regression: Save dialog box and OK in the Linear Regression dialog box, SPSS runs a regression analysis and saves the residuals as a new variable in your dataset. We will discuss all the output in detail below.

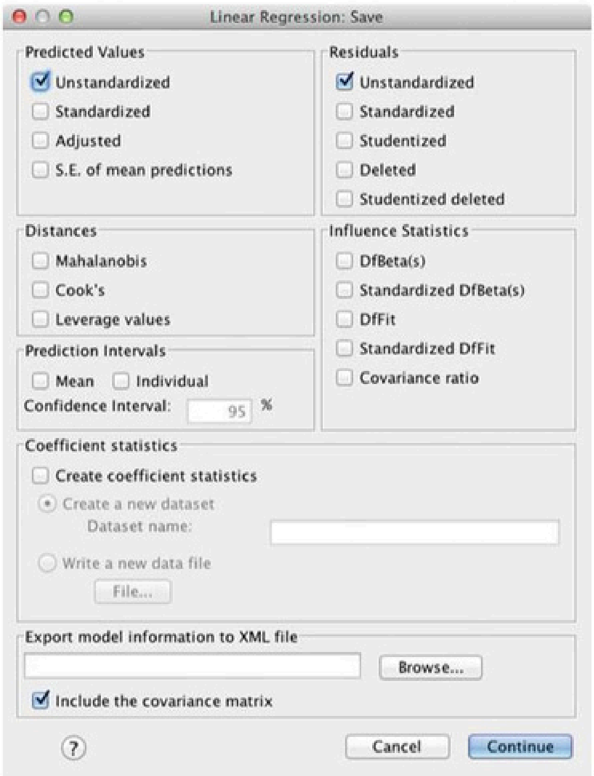


Fig. 7.11 The Save options for regression analysis

7.4.3 Test the Assumptions of Regression Analysis Using SPSS

To test the assumptions, we need to run three separate analyses.

The first assumption, the regression model can be expressed in a linear way, is implied if you can write the regression model linearly as y ¼ α þ β1x1 þ β2x2 þ β3x3 þ e. This is easy to do! While not needed, we also check for linearity of the relationships between the independent and dependent variables. If we create a scatter plot of Overall Customer Satisfaction against Customer Expectations, SPSS produces Fig. 7.12. This plot seems to suggest a linear relationship between the two variables. For a full analysis, we should plot every separate independent variable against the dependent variable. Try this yourself and you will see that the other independent variables are also linearly related to Overall Customer Satisfaction. Note that Perceived Value includes a clear outlier but with or without this outlier the relationship is still linear. Also note that if we include variables that take on only few different values, such as Customer Complaints (values of 0 and 1), we cannot use graphs to see if these relationships are linear.

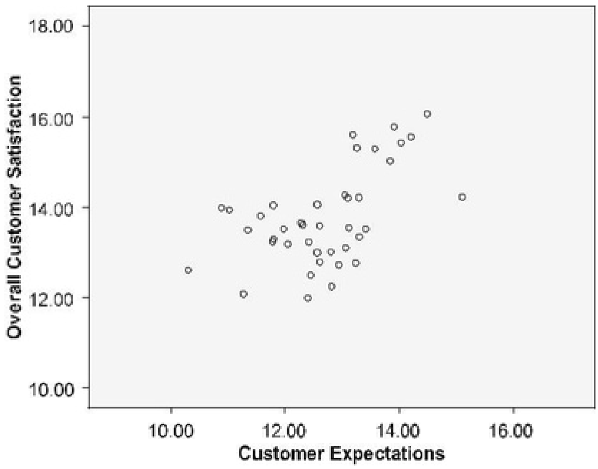


Fig. 7.12 Overall customer satisfaction against customer expectations

Next, we have to check if the regression model's expected mean error is zero (second assumption). Remember, this choice is made on theoretical grounds. We have a randomly drawn sample from the population and the model is similar in specification to other models explaining customer satisfaction. This makes it highly likely that the regression model's expected mean error is zero.

The third assumption is that of homoskedasticity. To test for this, we plot the errors against the dependent variable. Do this by going to u Graphs u Legacy Dialogs u Scatter/Dot (choose Simple Scatter). Enter the Overall Customer Satis-faction to the Y-axis and put Unstandardized Residual to the X-axis. Then click on OK. SPSS then produces a plot similar to Fig. 7.13. The results do not suggest heteroskedasticity. Note that it clearly seems that there is an outlier present. By looking at the Casewise Diagnostics in Table 7.4, we can further investigate this issue (note that we have set Casewise diagnostics to "Outliers outside: 3 standard deviations" to be conservative).

Cases where the errors are high indicate that those cases influence the results. SPSS assumes by default (see Fig. 7.10 under Outliers outside: 3 standard deviations) that observations that are three standard deviations away from the mean are potential outliers. The results in Table 7.4 suggest that there are four potential outliers. Case 257 has the strongest influence on the results (the standardized error is the highest). Should we delete these four cases? Case 257 seems to be very far away from the other observations (also see Fig. 7.13) and is likely an entry error or mistake, meaning the observation should be deleted. The other potential outliers appear to be simply part of the data, and should be retained.

If we delete a case from the initial dataset, we have to re-run the model. When doing so, we have to re-consider the assumptions we just discussed based on the newly estimated unstandardized residuals. However, we refrain from displaying the results twice – just try it yourself! Let's instead continue by discussing the remaining two (partly optional) assumptions using the dataset without the outlier.[7]

7You can download the reduced dataset ACSI Data_without outlier.sav in the 8 Web Appendix (Chap. 7)

Fig. 7.13 A plot of the errors against the dependent variable's values

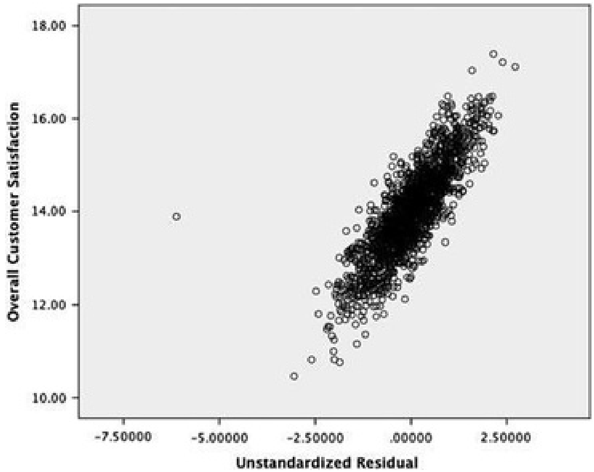


Table 7.4 Casewise diagnostics

Casewise Diagnosticsa

Case Number	Std. Residual	Overall Customer Satisfaction	Predicted Value	Residual
257	-7.212	13.89	20.0091	-6.12239
600	-3.594	10.46	13.5149	-3.05122
655	3.200	17.12	14.3997	2.71669
1044	-3.053	10.82	13.4133	-2.59206

a. Dependent Variable: Overall Customer Satisfaction

If we had data with a time component, we would also perform the Durbin– Watson test to test for potential autocorrelation (fourth assumption). However, since the data do not include any time component, we should not conduct this test.

Lastly, we should explore how the errors are distributed. Do this by going to u Graphs u Legacy Dialogs u Histogram. Enter the Unstandardized Residual under Variable. Also make sure that you check Display normal curve. In Fig. 7.14, we show a histogram of the errors.

Figure 7.13 suggest that our data are normally distributed as the bars indicating the frequency of the errors generally follow the normal curve. However, we can check this further by conducting the Kolmogorov-Smirnov test (with Lilliefors correction) by goingto Analyze Descriptive Statistics Explore. Table 7.5 shows the output.

The results of this test (Table 7.5) suggest that the errors are normally distributed as we do not reject the test's null hypothesis. Thus, we can assume that the errors are normally distributed.

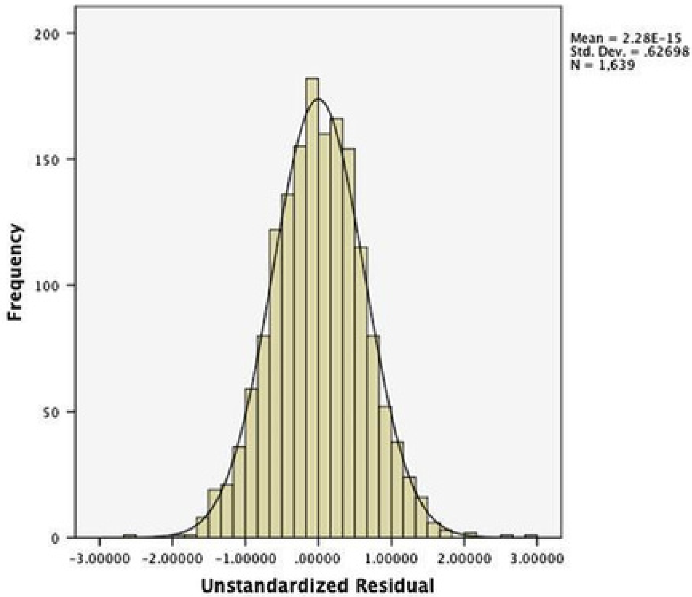


Fig. 7.14 Histogram of the errors with a standard normal curve

Table 7.5 Output produced by the Kolmogorov–Smirnov test

Tests of Normality

	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
UnstandardizedResidual	.016	1639	.200*	.998	1639	.041

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Now we turn to testing for multicollinearity. There are two tables in which SPSS indicates if multicollinearity issues are present. The first table is Table 7.11, in which the regression coefficient estimates are displayed. This table output also shows each variable's Tolerance in the second to last column and VIF in the last column. In this example, the tolerance values clearly lie above 0.10, and VIF values below 10, indicating that multicollinearity is of no concern.

Table 7.6 Descriptive statistics table

Descriptive Statistics

	Mean	Std. Deviation	N
Overall Customer Satisfaction	13.9999	1.00219	1639
Customer Expectations	13.0009	.99957	1639
Perceived Value	9.9990	1.01018	1639
Customer Complaints	.2288	.42019	1639

7.4.4 Interpret the Regression Model Using SPSS

The results of the regression analysis that we just carried out are presented below. We will discuss each element of the output that SPSS created in detail.

Tables 7.6 and 7.7 describe the dependent and independent variables in detail and provide several descriptives discussed in Chap. 5. Notice that the deletion of the outlier reduced the overall number of observations from 1,640 to 1,639. These are the observations for which we have complete information for the dependent and independent variables. Table 7.7 shows the correlation matrix and gives an idea how the different variables are related to each other.

SPSS also produces Table 7.8, which indicates the variables used as dependent and independent variables and how they were entered in the model. It confirms that we use the Enter option (indicated under Method). All independent variables included in the model are mentioned under Variables Entered and under b. Further-more, under Dependent Variable, the name of the dependent variable is indicated.

We interpret Tables 7.9 and 7.10 jointly, as they provide information on the model fit; that is, how well the independent variables relate to the dependent variable. The R2 provided in Table 7.9 seems satisfactory and is above the value of 0.30 that is common for cross-sectional research. Usually, as is the case in our analysis, the R2 and adjusted R2 are similar. If the adjusted R2 is substantially lower, this could indicate that you have used too many independent variables and that some could possibly be removed. Next, consider the significance of the F-test. The result in Table 7.10 indicates that the regression model is significant (Sig. <0.5).

After assessing the overall model fit, it is time to look at the individual coefficients. We find these in Table 7.11. First, you should look at the individual parameters' t-values, which test if the regression coefficients are individually equal to zero. If this is the case, the parameter is insignificant. In the model above, we find

Table 7.7 Correlation matrix

Correlations

		Over all Custo m er S atisfa ction	Custo m er Expectations	Perceived Value	Custo m er Com plain ts
Pearson Correlation	Overall Customer S atisfa ctio n	1.00 0	.492	.766	−.14 4
	Customer Expectations	.492	1.00 0	.478	−.07 3
	Perceived Value	.766	.478	1.0 00	−.13 7
	Customer Complaints	−.14 4	−.07 3	− .137	1.00 0
Sig.(1-tailed)	OverallCustomer S atisfa ctio n	.	.000	.000	.000
	Customer Expectations	.000	.	.000	.001
	Perceived Value	.000	.000	.	.000
	Customer Complaints	.000	.001	.000	.
N	OverallCustomer S atisfa ctio n	1639	1639	1639	1639
	Customer Expectations	1639	1639	1639	1639
	Perceived Value	1639	1639	1639	1639
	Customer Complaints	1639	1639	1639	1639

three significant coefficients, those with p-values (under Sig. in Table 7.11) are below the commonly used level of 0.05. Although the constant is also significant, this is not a variable and is usually excluded from further interpretation. The significant variables require further interpretation.

First look at the sign (plus or minus) in the Standardized Coefficients column. Here, we find that Customer Expectations and Perceived Value are significantly and positively related to Overall Customer Satisfaction. Customer Complaints is signif-icant and negatively related to Overall Customer Satisfaction. This means that if people complain, their customer satisfaction is significantly lower on average. By looking at the standardized coefficients' values you can assess if Customer Expectations, Perceived Value, or Customer Complaints is most strongly related to Overall Customer Satisfaction. You only look at the absolute value (without the minus or plus sign therefore) and choose the highest value. In this case, Perceived Value (0.677) has clearly the strongest relationship with overall customer satisfac-tion. Therefore, this might be the first variable you want to focus on through marketing activities if you aim to increase customer satisfaction. Although standardized βs cannot be fully compared when an independent variable is binary (as it is the case with Customer Complaints) the comparison gives us a rough idea regarding the relative strengths of the independent variables' effects on the depen-dent variable.

Table 7.8 Variables used and regression method

Variables Entered/Removedb

Model	Variables Entered	Variables Removed	Method
1	Customer Complaints, Customer Expectations, Perceived Valuea	.	Enter

a. All requested variables entered.
b. Dependent Variable: Overall Customer Satisfaction

Table 7.9 The model summarya

Model Summaryb

Model	R	RSquare	Adjusted R Square	Std. Error of the Estimate
1	.780a	.609	.608	.62756

a. Predictors: (Constant), Customer Complaints, Customer Expectations, Perceived Value
b. Dependent Variable: Overall Customer Satisfaction

Table 7.10 ANOVA a

ANOVAb

Model		SumofSquares	df	MeanSquare	F	Sig.
1	Regression	1001.261	3	333.754	847.453	.000a
	Residual	643.914	1635	.394		
	Total	1645.175	1638			

a. Predictors: (Constant), Customer Complaints, Customer Expectations, Perceived Value
b. Dependent Variable: Overall Customer Satisfaction

Table 7.11 The estimated coefficients

Coe ffic i e ntsa

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std.Error	Beta			Tolerance	VIF
1	(Constant)	5.124	.215		23.863	.000		
	CustomerExpectations	.164	.018	.163	9.257	.000	.77	1.296
	PerceivedValue	.677	.018	.683	38.484	.000	1	1.31
	CustomerComplaints	−.092	.037	−.03	−2.458	.014	.76	4
				8			1	1.01
							.98	9
							1	

a. Dependent Variable: Overall Customer Satisfaction

The Unstandardized Coefficients column gives you an indication of what would happen if you were to increase one of the independent variables by exactly one unit. For example, if Customer Expectations were to increase by one unit, we would expect Overall Customer Satisfaction to increase by 0.164 units. The stan-dard errors are used to calculate the t-values. If we take the unstandardized coeffi-cient of Customer Expectations (0.164) and divide this by its standard error (0.018), we obtain a value that is approximately the t-value of the 9.257 indicated in the table (see Chap. 6 for a description of the t-test statistic). The slight differences are due to rounding. As indicated before, Customer Complaints is a binary variable which can only take the value of 0 or 1. More precisely, for those customers who have not complained thus far, Customer Complaints takes the value 0. On the contrary, if a customer has already complained, the variable's value is 1 for this observation. Thus, the corresponding coefficient (-0.092) is the difference in satis-faction for customers who complained compared to those who have not complained. Overall, the results indicate that we have found a useful model that satisfies the assumptions of regression
analys is.

7.4.5 Validate the Regression Model Using SPSS

Next, we need to validate the model. Let's first split-validate our model. Do this by going to u Data u Select Cases. This displays a dialog box similar to Fig. 7.15.

In this dialog box, go to Select Cases: Range. This displays a dialog box similar to Fig. 7.16.

Select the first 1,150 cases, which is approximately 70% of the data. Then run the regression analysis again. Afterwards, return to Select Cases: Range and select observations 1,151–1,639. Compare the results of this model to those of the previous model. This approach is simple to execute but only works if the ordering of the data are random.

Next, we can add a few key additional variables to our model and see if the basic results change. Key variables with which to check the stability (the so-called covariates) could be the total annual family income and the respondent's gender. Then interpret the basic model again to see if the regression results change.

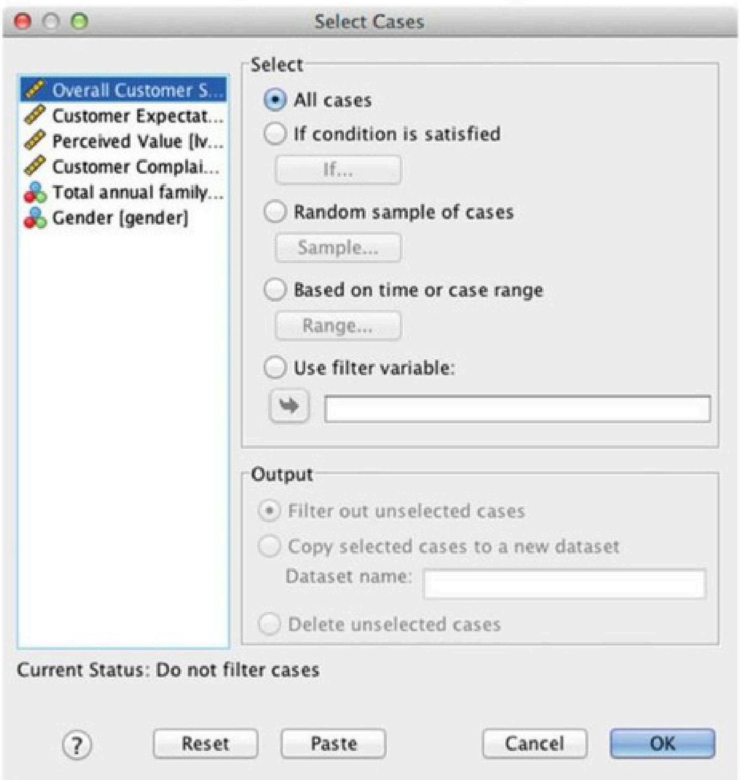


Fig. 7.15 The select cases dialog box

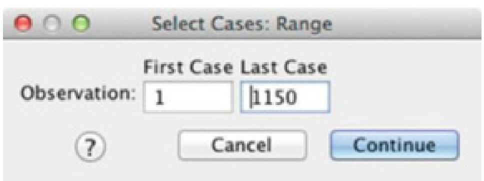


Fig. 7.16 The select cases: range dialog box

7.5 Farming with AgriPro (Case Study)

AgriPro is a firm based in Colorado, USA, which does research on and produces genetically modified wheat seed. Every year AgriPro conducts thousands of experiments on different varieties of wheat seeds in different locations of the USA. In these experiments, the agricultural and economic characteristics, regional adaptation, and yield potential of different varieties of wheat seeds are investigated. In addition, the benefits of the wheat produced, including the milling and baking quality, are examined. If a new variety of wheat seed with superior characteristics is identified, AgriPro produces and markets it throughout the USA and parts of Canada.

AgriPro's product is sold to farmers through their distributors, known in the industry as growers. Growers buy wheat seed from AgriPro, grow wheat, harvest the seeds, and sell the seed to local farmers, who plant them in their fields. These growers also provide the farmers who buy their seeds with expert local knowledge on management and the environment.

AgriPro sells its products to these growers in several geographically defined markets. These markets are geographically defined because different local conditions (soil, weather, and local plant diseases) force AgriPro to produce different products. One of these markets, the heartland region of the USA is an important market for AgriPro, but the company has been performing below management expectations in these markets. The heartland region includes the states of Ohio, Indiana, Missouri, Illinois, and Kentucky.

To help AgriPro understand more about farming in the heartland region, they commissioned a marketing research project among farmers in these states. AgriPro, together with a marketing research firm, designed a survey, which included questions on what farmers who decide to plant wheat find important, how they obtain information on growing and planting wheat, what is important in their purchasing decision, and their loyalty to and satisfaction with the top five wheat suppliers (including AgriPro). In addition, questions were asked about how many acres of farmland the respondents possessed, how much wheat they planted, how old they were, and their level of education.

http://www.agriprowheat.com

This survey was mailed to 650 farmers selected from a commercial list that includes nearly all farmers in the heartland region. In all, 150 responses were received, resulting in a 23% response rate. The marketing research firm also assisted AgriPro to assign variable names and labels. They did not delete any questions or observations due to nonresponse to items.

Your task is to analyze the dataset further and provide the management of AgriPro with advice based on the dataset. This dataset is labeled Agripro.sav and is available in the 8 Web Appendix (! Chap. 7). Note that the dataset (under

Variable View at the bottom of the SPSS screen) contains the variable names and labels and these match those in the survey. In the 8 Web Appendix (! Chap. 7), we also include the original survey.8

To help you with this task, a number of questions have been prepared by AgriPro that they would like to see answered:

1. Produce appropriate descriptive statistics for each item in the dataset. Consider descriptive statistics that provide useful information in a succinct way. In addition, produce several descriptive statistics on the demographic variables in the dataset, using appropriate charts and/or graphs.
2. Are there any outliers in the data? What (if any) observations do you consider to be outliers and what would you do with these?
3. What are the most common reasons for farmers to plant wheat? From which source are farmers most likely to seek information on wheat? Is this source also the most reliable one?
4. Consider the five brands included in the dataset. Describe how these brands compare on quality, advice provided, and farmer loyalty.
5. How satisfied are the farmers with the brand's distributors?
6. AgriPro expects that farmers who are more satisfied with their products devote a greater percentage of the total number of acres available to them to wheat. Please test this assumption by using regression analysis. In addition, check the assumptions of regression analysis.
7. Is there a relationship between farmers' satisfaction with AgriPro and the respondent's educational level, age, and number of acres of farmland? Conduct a regression analysis with all these four variables. How do these results relate to question 6?
8. Are all assumptions satisfied? If not, is there anything we can do about it or should we ignore the assumptions if they are not satisfied?
9. What is the relationship between the quality of AgriPro seed and the satisfaction with AgriPro?
10. As AgriPro's consultant, and based on the empirical findings of this study, what marketing advice would you have for AgriPro's marketing team? Provide four or five carefully thought through suggestions as bullet points.