# Chapter 7
# 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** Regression analysis · Errors · Residuals · Ordinary least squares · $R^2$ · Adjusted $R^2$ · F-test · Sample size · Linearity · Outliers · Heteroskedasticity · (Multi)collinearity · Autocorrelation · Durbin-Watson test · Kolmogorov-Smirnov test · Shapiro-Wilk test

Many organizations, including Procter & Gamble, McKinsey & Company, and Nestlé use regression analysis to help make key marketing decisions. They can, for example, support decisions on pricing, trade promotions, advertising, and distribution with regression analysis. Regression analysis helps these organizations by providing precise quantitative information on which managers can base their decisions. In this chapter, we explain how you can use regression analysis for market research.

## 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. Mooi and M. Sarstedt, *A Concise Guide to Market Research*,
DOI 10.1007/978-3-642-12541-6_7, 

(e.g., 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 if sales depend more strongly on price or on advertising. Most importantly, regression analysis allows us to compare the effects of variables measured on different scales such as the effect of price changes (e.g., measured in USD) and the number of promotional activities.

Regression analysis can also help make predictions. For example, if we have data on sales, prices, and promotional activities, 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.

## 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? Regression analysis is a way of fitting a "best" line through a series of observations. Figure 7.1 plots a dependent ($y$) variable (weekly supermarket sales in USD) against an independent ($x$) variable (an index of promotional activities). With "best" line we mean that it is fitted in such a way that it results in the lowest 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 in a certain way.

Before we start explaining regression analysis further, we should discuss regression notation. Regression models are generally described 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

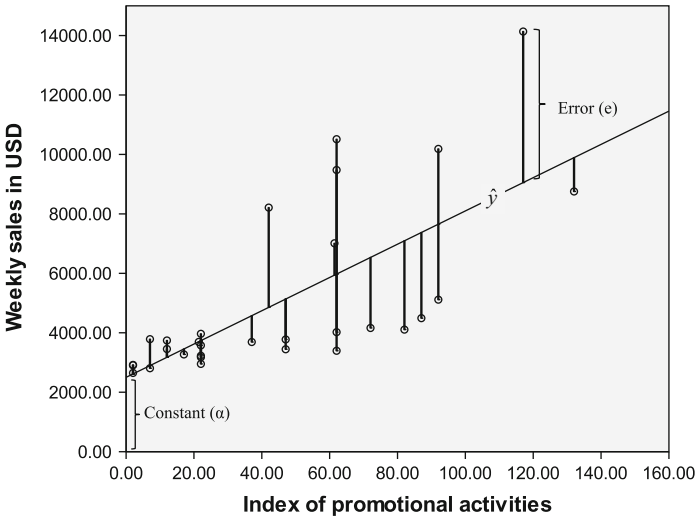


**Fig. 7.1** A visual explanation of regression analysis

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 USD. 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 is always included to make sure that the regression model has the best possible fit with the data.

The independent variable is indicated by $x_1$. As we have only one independent variable, we only see one $x$ in our model. $\beta_1$ (pronounced as *beta*) indicates the (regression) coefficient of the independent variable. This coefficient represents the gradient of the line and is also referred to as the *slope*. A positive $\beta_1$ coefficient indicates an upward sloping regression line while a negative $\beta_1$ 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 $\beta_1$ as 55.968, meaning that if we increase promotional activities by one unit, sales will go up by 55.968 USD on average. In regression analysis, we can calculate whether this value (the $\beta_1$ parameter) differs significantly from zero by using t-tests.

The last element of the notation, the $e$ denotes the *error* or *residual* of the equation. The term error is commonly used in research. However, SPSS refers to errors as 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 (or residual) 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., $\beta_1$, $\beta_2$, and $\beta_3$). 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 examine 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. In this step, we discuss the basics, such as which independent variables to select. Thereafter, we discuss the assumptions of regression analysis. Next, we interpret and validate the model. The last step is to use the regression model to make predictions.

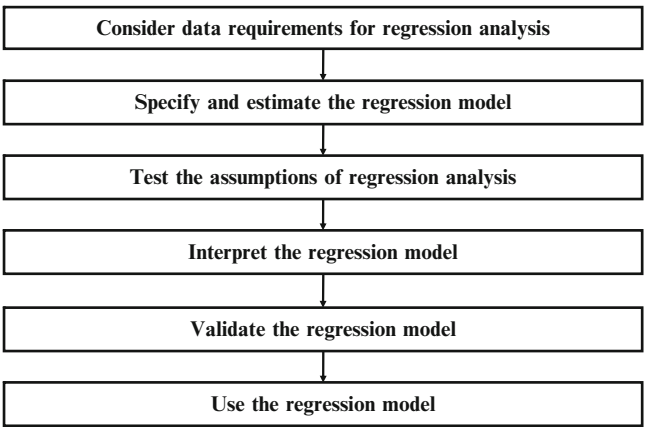


**Fig. 7.2** Steps to conduct a regression analysis.

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

[2]This only applies to the standardized βs.

# Conducting a Regression Analysis

## *Consider Data Requirements for Regression Analysis*

Several data requirements have to be considered before we undertake a regression analysis.

A 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 approach is the formal one in which we use power analysis to calculate the minimum sample size required to accept the regression outcomes with a particular level of confidence. As mentioned in Chap. 6 (Box 6.2), these calculations are difficult. Fortunately, the Internet offers a wide selection of downloadable applications and interactive Web programs that will simplify your work. Just google "power analysis calculator regression" and you will find numerous helpful sites. Most require you to specify several parameters, such as the number of independent variables or the sample size, to calculate the resulting level of power. By convention, 0.80 is an acceptable level of power. Kelley and Maxwell (2003) also discuss sample size requirements in this context.

The second approach is through rules of thumb. These rules are not specific to a situation and 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 the overall relationships between the independent and dependent variable, the number of observations is at least $50 + 8 \cdot k$ (where $k$ are the number of independent variables). Thus, for a model with ten independent variables, you need $50 + 8 \times 10 = 130$ observations, or more. If you want to test for individual parameters' effect (i.e., if one coefficient is significant or not), Green proposes a sample size of $104 + k$. Thus, if you have ten independent variables, you need $104 + 10 = 114$ observations. If you are interested in both the general relationship, and the individual coefficients, you should take the higher number of required observations of the two formulas.[3]

A related data requirement is that the sample used is representative of the population (see Chap. 3 for a discussion on samples and populations). Imagine that we want to understand what drives the satisfaction of US car owners. If our sample is collected just from executives (who typically earn high incomes), our results may be biased, as high-income consumers usually own more expensive cars, such as a BMW, Lexus, or Mercedes, than the total population of US car owners. If

[3]Rules of thumb are almost never without drawbacks and caveats. For Green's formula, these are that you need a larger sample size than he proposes if you expect small effects (thus a low expected $R^2$ such as 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 $R^2$ is low, and about twice the required sample size in case of measurement errors or if stepwise methods are used.

we run regression analysis using such data, our results reflect those for executives, and not those of the population.

A third data requirement is that no regression model works if the variables have no variation. Specifically, if there is no variation in the dependent variable (i.e., it is a constant), we also do not need regression, as we already know what the dependent variable's outcome is. Likewise, if an independent variable has no variation, it cannot explain any variation in the dependent variable.

A fourth regression analysis data requirement is that the dependent variable needs to be interval or ratio scaled. To determine if data are interval or ratio scaled, consult 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 nominal variable with more than two levels, you can use multinomial logistic regression although OLS is also used for this purpose in practice. This should, for example, be used if you want to explain why people prefer product A over B or C. Finally, if the dependent variable is ordinally scaled (e.g., to predict first, second, and third choice) you should use ordinal regression. We do not discuss these different methods in this chapter, but they are intuitively similar to regression. For a comprehensible discussion of regression methods with dependent variables measured on a nominal or ordinal scale, see Field (2009).

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. If more than two independent variables are highly correlated, we talk about *multicollinearity*. *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). This may happen because you entered the same independent variable twice, or because one variable is a linear combination of the other (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 satisfied customers are with 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 these weaker forms of collinearity is that they tend to decrease the significance of independent variables.

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

[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 $R^2$ 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.

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. If weaker forms of collinearity occur, it is up to you to decide what to do. Firstly, if you have multiple overlapping variables, you could conduct a factor analysis first (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. Alternatively, you could 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. Another good way of identifying variables that cause multicollinearity is by looking at the condition index output in SPSS. Since this output is quite advanced, we refrain from discussing it in greater detail here. On Garson's Statnotes regression page (Box 7.1) you will find a brief explanation of this additional table.

**Box 7.1** Advanced collinearity diagnostics

http://faculty.chass.ncsu.edu/garson/PA765/regressa.htm

## *Specify and Estimate the Regression Model*

To conduct a regression analysis, we need to select the variables we want to include, decide on how they are included, and decide on how the model is estimated. 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 *Sales data.sav* (🖱 Web Appendix → Chap. 7). These data contain information on

supermarket sales per week in USD (*Sales*), the (average) price level (*Price*), and an index of promotional activities (*Promotion*), amongst other variables. After having opened the dataset, click on ▶ Analyze ▶ Regression ▶ Linear. This opens a box similar to Fig. 7.3.

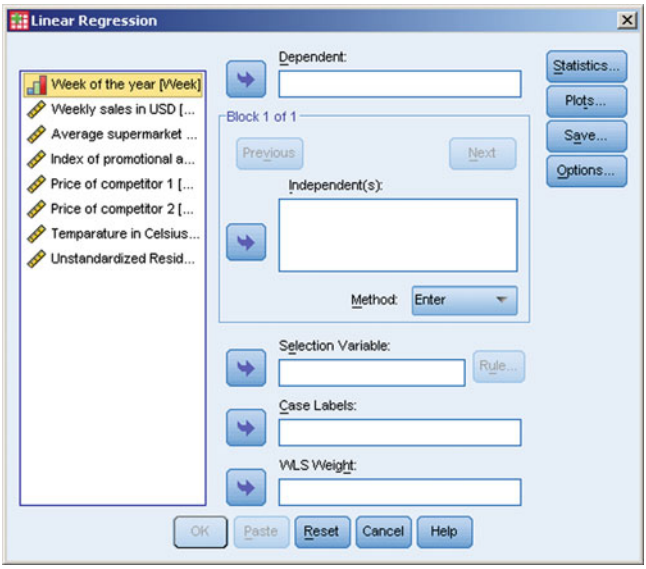


**Fig. 7.3** The main regression dialog box in SPSS

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 supermarket 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. 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.

For this example, we use *sales* as the dependent variable and *price* as well as the *index of promotional activities* as independent variables.

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.

Starting with the constant ($\alpha$) 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.

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 complete information is present. Generally, backward or forward models result in finding highly significant models that only use a small number of observations from the total number of available observations. In this case, a regression model results that fits a small set of the data well but not the whole data set 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 calculated. 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 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 distances of regression points above and below the line, the result is exactly zero. In certain situations, OLS, does not work very well and we need to use alternative procedures such as *Weighted Least Squares (WLS)*. We briefly discuss when WLS should be used. Greene's (2007) work on WLS is a good source of further information. Greene also discusses other estimation procedures such as *Two-staged least squares* that are beyond the scope of this book.

A last point 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 choose to 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 regression 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 regression models that try to explain economic growth, regression models can have dozens of independent variables.

## *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 regression analysis assumptions. 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:

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

A fifth assumption is optional. If we meet this assumption, we have information 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.

- The errors need to be approximately normally distributed.

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

The first assumption means that we can write the regression model as $y = \alpha + \beta_1 x_1 + e$. Thus, relationships such as ${\beta_1}^2 x_1$ 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 $\beta$s) are linear, you satisfy this assumption. You can, however, transform the $x$ variables any way that is necessary to produce a best fitting regression line. For example, if you take the log of the $x$ variable, the relationship between the $y$ and $x$ variables becomes non-linear, but you still satisfy this assumption because the $\beta$ is linear. In practice, most relationships between $x$ and $y$ variables are linear but there are also many exceptions, sometimes quite dramatic ones.

Checking the linearity between $y$ and $x$ variables can be done by plotting the independent variables against the dependent variable. Using this 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 a dashed, upward sloping, and another 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 relationship, because if we specify a relationship as linear when it is in fact non-linear, the regression analysis results will be biased.

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 USD 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

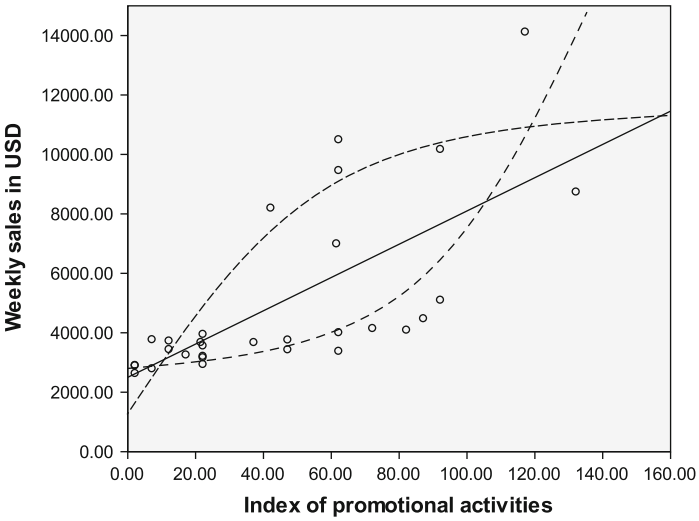


**Fig. 7.4** Different relationships between promotional activities and weekly sales

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). Furthermore, omitting important independent variables could cause the expected mean not to be zero. Simply put, if we were to omit a relevant variable $x_2$ from a regression model that only includes $x_1$, we induce a bias in the regression model. More precisely, $\beta_1$ is likely to be inflated, which means that the estimated value is higher than it should actually be. Thus, $\beta_1$ itself is biased because we omit $x_2$!

The third assumption is that the errors' variance is constant, a situation we frequently call *homoskedasticity*. Imagine that we want to explain the weekly sales of various supermarkets in USD. Clearly, large stores have a much larger spread in sales than small supermarkets. For example, if you have average weekly sales of 50,000 USD, you might see a sudden jump to 60,000 USD or a fall to 40,000 USD. However, a very large supermarket could see sales move from an average of 5,000,000–7,000,000 USD. 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 $\beta$s 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.

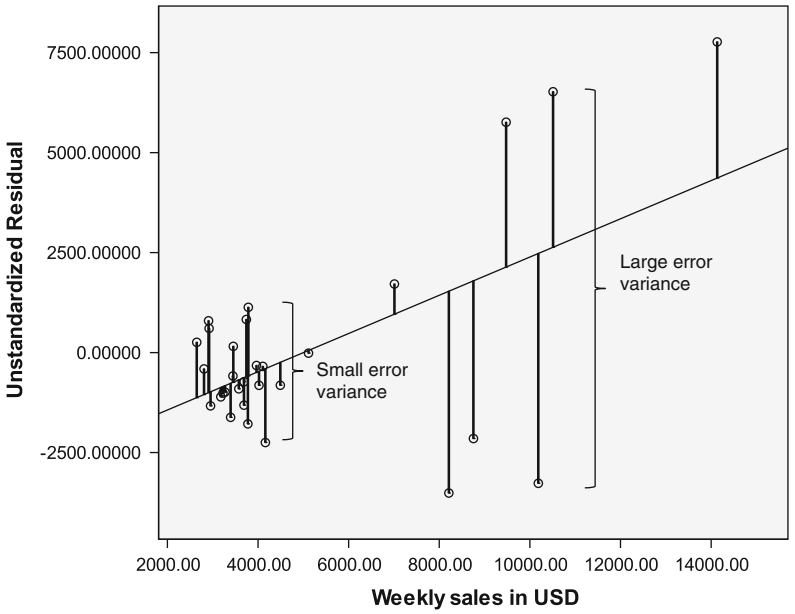


**Fig. 7.5** The heteroskedasticity issue

On theoretical grounds, try to understand whether it is likely that the errors increase as the value of the dependent variable in(de)creases. 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 in(de)creases, the variance should not in(de)crease. 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 variance not to be constant (e.g., store size) and "weight" the results by this variable. Only use WLS if 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.

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. If this is rejected, 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. 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 at different points in time. Do this by going to ▶ Data ▶ 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 three 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). If, on the other hand, 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. If no systematic pattern of errors occurs, we have no autocorrelation. These Durbin–Watson values are used to test a null hypothesis of no autocorrelation. Depending on the calculated Durbin–Watson test statistic values, we can draw three different conclusions:

- The calculated Durbin–Watson value falls below the lower critical value. In this case, we have negative autocorrelation.
- The calculated Durbin–Watson value falls above the upper critical value. In this case, we have positive autocorrelation.
- The calculated Durbin–Watson value falls in between the lower and higher critical value. In this case, we have no autocorrelation.

The critical values can be found on the website accompanying this book (🖱 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. 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 models. We do not discuss these in this book, but a useful source of further information is Hill et al. (2008).

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 regression model still provides good estimates of the coefficients, particularly if the

sample size used is reasonably large (above the recommended minimum sample size discussed previously). Therefore, we consider this assumption as optional.

Potential reasons for non-normality include outliers (discussed in Box 7.2) and a non-linear relationship between an independent and a dependent variable.

**Box 7.2** Outliers

Outliers can influence regression results dramatically. When running regression analysis, it is important to be aware of outliers in the data. Study Fig. 7.6. The observation in the upper left corner (indicated with the solid circle) is probably an outlier as it differs very much from all the other observations. Without the outlier, the slope of the regression line (indicated by the dashed line) is positive. If we include the outlier, the slope becomes negative. Including the outlier produces an effect that is opposite to the results without it!

We can detect outliers in several ways. We can try to detect outliers by looking at one variable at a time, by looking at the relationship between two variables, or by looking at a regression model's errors. Next, we discuss several ways to identify outliers, which should always be used as each method has its advantages and disadvantages.

If we investigate one variable at a time, we can calculate the minimum and maximum of each variable and see if unexpectedly high or low values occur. If the minimum or maximum falls outside the expected range, for example, because we have made a typing error, we can identify that observation and subsequently run a regression model without that observation. Researchers often consider values more than three standard deviations away from the mean as potential outliers.

Scatter plots are another good way to identify outliers that are the result of a combination of extreme values for two variables. For example, if we look at Fig. 7.6, we can see that the observation in the upper right corner is very different in that it is both very high on the y-axis and on the x-axis. These kinds of observations can strongly influence the results, much more so than if the outlier only had very high $x$ or $y$ values.

We can also check if the regression errors of some observations are very large. This is particularly important if the regression model has multiple independent variables. SPSS can do this using **Casewise Diagnostics**, which is included in the **Statistics** option included in the main regression dialog box. Casewise diagnostics indicate if large errors occur. Remember that errors are very large if the particular observation is very far away from the regression line. If a particular observation's error is large, that observation "pulls" the regression line upwards (if the error is positive) or downwards (if the error is negative). Regression errors that occur at extreme values for the independent variable are particularly influential, because these have a great deal of "leverage," meaning they may shift the entire line up (or down) and change the slope of the line (from

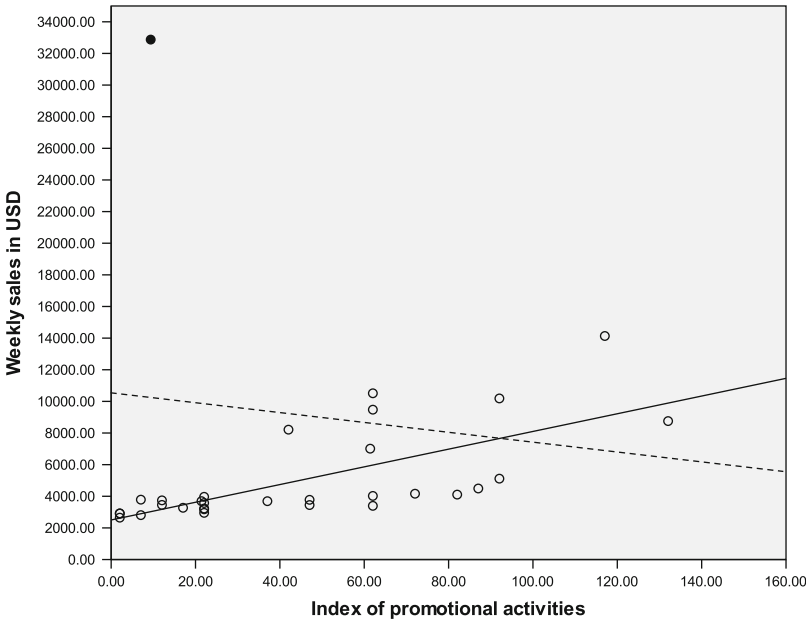


**Fig. 7.6** Outliers

positive to negative and vice versa). In Chap. 5, we further discuss how to identify outliers and how to deal with them.

The critical decision with outliers is whether they should be retained, transformed, or deleted. This decision is somewhat arbitrary as there are no fixed recommendations on what to do. The only fixed recommendation is that if outliers are deleted, you should clearly indicate why you have chosen to do so.

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

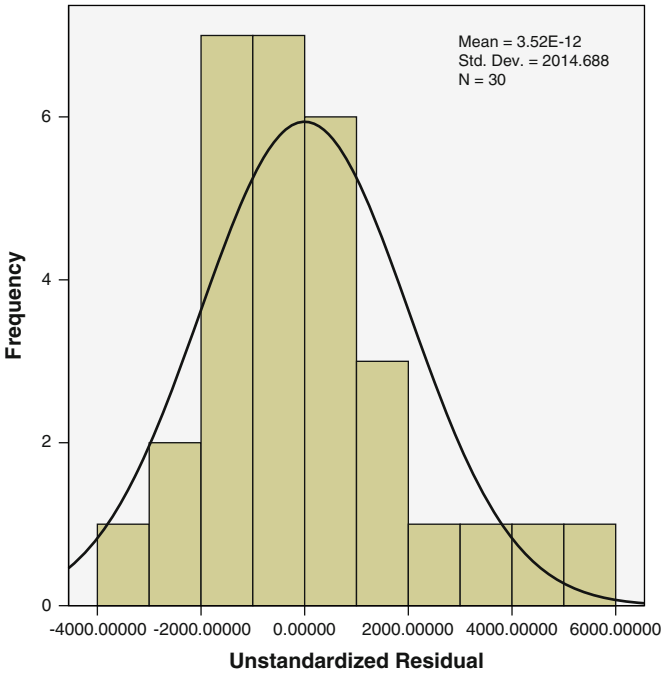


**Fig. 7.7** Histogram of the errors

**Table 7.1** Output produced by the Shapiro–Wilk test

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Unstandardized Residual | .131 | 30 | .200* | .939 | 30 | .084 |

a. Lilliefors Significance Correction

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

Since we have less than 50 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 = 0.084) (Table 7.1).

## *Interpret the Regression Model*

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.

## *Overall Model Fit*

We can assess the overall model fit using the (adjusted) $R^2$ and significance of the F-value.

The $R^2$ (or *coefficient of determination*) indicates the degree to which the model explains the observed variation in the dependent variable. In Fig. 7.8, we explain this graphically with a scatter plot. The y-axis relates to the dependent variable (weekly sales in USD) 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 USD 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 $SS_T$). We indicate the total variation for only one observation on the right of the scatter plot.

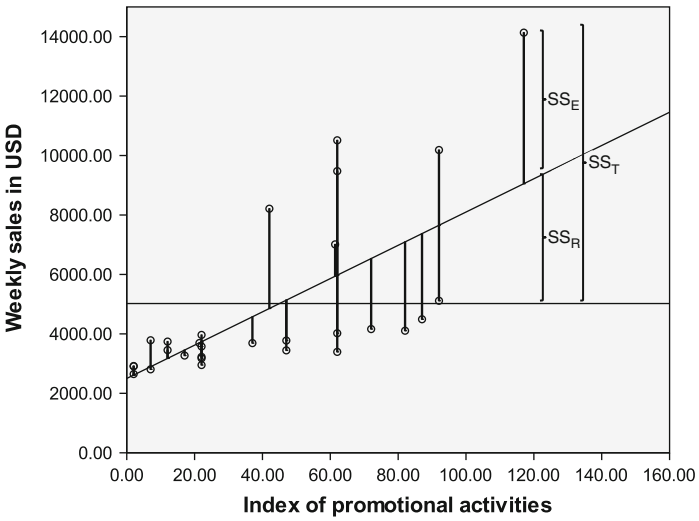


**Fig. 7.8** Explanation of the $R^2$

The upward sloping line (starting at the y-axis at about 2,500 USD sales per week) is the regression line that is estimated 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 $SS_R$) 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 $SS_E$. In more formal terms, we can describe these types of variation as follows:

$$SS_T = SS_R + SS_E$$

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 $\bar{y}$ 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 $SS_R$) relative to the total variation ($SS_T$). This is the $R^2$ and we can calculate this as:

$$R^2 = \frac{SS_R}{SS_T}$$

The $R^2$ always lies between 0 and 1, where a higher $R^2$ indicates a better model fit. When interpreting the $R^2$, higher values indicate that more of the variation in $y$ is explained by variation in $x$, and therefore that the $SS_E$ is low. The $R^2$ is the total relationship between the independent variables and the dependent variable. It is difficult to provide rules of thumb regarding what $R^2$ is appropriate, as this varies from research area to research area. For example, in longitudinal studies $R^2$s 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.

If we use the $R^2$ 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 $R^2$ will increase. Thus, if we use the $R^2$ 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 $R^2$ is plainly not a good strategy, as we want regression models that do a good job of explaining the data (thus a low $SS_E$), but which also have few independent variables (these are called *parsimonious models*). We do not want too many independent variables because this makes the regression model results less actionable. 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* $R^2$ to select regression models. The adjusted $R^2$ only increases if the addition of another

independent variable explains a substantial amount of variance. We calculate the adjusted $R^2$ as follows:

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

Here, $n$ describes the number of observations and $k$ the number of independent variables (not counting the constant $\alpha$). This adjusted $R^2$ 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 $R^2$ when comparing regression models.

However, do not blindly use the adjusted $R^2$ 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 $\beta$ differs from zero. Thus, if we reject this null hypothesis, we conclude that the model is most likely useful. However, 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:

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

The F-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, when considered in isolation, significant. This is due to testing differences – the F-test is used for the entire model, and the t-test for individual parameters – and due to issues such as multicollinearity.

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.

## *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 has a significant influence on the dependent variable

To be precise, the null and alternative hypotheses tested for an individual parameter (e.g., $\beta_1$) are:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

If a coefficient is significant (meaning we reject the null hypothesis), we also need to look at the unstandardized $\beta$ coefficient and the standardized $\beta$ coefficient. The unstandardized $\beta$ 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 USD with a $\beta_1$ coefficient of 55.968. This means that if we increase promotional activities by one unit, weekly sales will go up by 55.968 USD. In other cases, the effect could of course also be negative (e.g., increasing prices reduces sales). Importantly, if we have multiple independent variables, the unstandardized $\beta_1$ 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 USD, 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) $\beta$ coefficients would in any case amount to comparing apples with oranges!

Fortunately, the standardized $\beta s$ allow us to compare the relative effect of differently measured independent variables. This is achieved by expressing $\beta$ as standard deviations with a mean of zero. The standardized $\beta s$ 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 $\beta s$ in this respect, as insignificant $\beta s$ do not (statistically) differ from zero! Practically, the standardized $\beta$ 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.

However, while the standardized $\beta s$ are helpful from a practical point of view, there are two problems associated with their usage. First, standardized β allow comparing the coefficients only within and not between models! Even if you add just a single variable to your regression model, standardized $\beta s$ may change substantially. Second, standardized $\beta s$ are not meaningful when the independent variable is a binary variable. Standardized $\beta s$ provide insight into the unique (as opposed to total) relationship with the dependent variable. In certain situations, however, it may seem that one independent variable has a slight relationship with the dependent variable (low standardized β), even if the actual (total) effects are strong. This happens if this variable exhibits high correlations with other independent variables (multicollinearity) which exert a stronger influence on the dependent than itself.

When interpreting (standardized) β coefficients, you should always keep the effect size in mind. If a β coefficient is significant, it indicates some effect that differs from zero. However, this does not necessarily mean that it is managerially relevant. For example, we may find that increasing promotional activities' unstandardized effect on sales in USD is just 0.01, yet statistically significant. Statistically we could conclude that the effect of a single increase in promotional activities increases sales by an average of 0.01 USD (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. An interesting perspective on significance and effect sizes is offered by Cohen's (1994) classical article "The Earth is Round ($p < .05$)."

## *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.

First, 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.

Second, 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.

Lastly, 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 assortment 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 $p = 0.051$).

Table 7.2 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 7.2** Key steps involved in carrying out a regression analysis

| Theory | Execution in SPSS |
|---|---|
| *Issues with regression analysis* | |
| Is the sample size sufficient? | Check if sample size is $50 + 8$ k or $104 + $ k, where $k$ indicates the number of independent variables. |
| Is the sample representative? | Check the recommendations made in Chap. 3. |
| Do the dependent and independent variables show variation? | Calculate the standard deviation of the variables by going to ▶ Analyze ▶ Descriptive Statistics ▶ Descriptives ▶ Options (check **Std. Deviation**). At the very least, the standard deviation should be a positive value. |
| Is the dependent variable interval or ratio scaled? | Use Chap. 2 to determine the measurement level. |
| Is (multi)collinearity present? | Check for tolerance and VIF. Do this with ▶ Analyze ▶ Regression ▶ Linear ▶ Statistics (check **Collinearity diagnostics**). The tolerance should be above 0.10. 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 |

(*continued*)

**Table 7.2** (continued)

| Theory | Execution in SPSS |
|---|---|
| | variables that could have a relationship with the dependent variable. Do not add all variables regardless. |
| *Testing the assumptions of regression analysis* | |
| Is the relationship between the independent and dependent variables linear? | Consider whether you can write the regression model as $y = \alpha + \beta_1 x_1 + \ldots + e$.<br>To understand if the independent variables are linearly related to the dependent variable, plot the *y* variables separately against the dependent variable of the regression model. Create scatter plots using ▶ Graphs ▶ Legacy Dialogs ▶ 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 (no heteroskedasticity)? | Plot the residual of the regression model on the y-axis and the dependent variable on the x-axis, using a scatter plot by means of ▶ Graphs ▶ Legacy Dialogs ▶ 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. |
| Test whether the errors are not correlated with one another | 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 value with the critical values. If autocorrelation is present and panel or time-series models need to be used:<br>▶ Analyze ▶ Regression ▶ Linear ▶ 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: ▶ Graphs ▶ Legacy Dialogs ▶ Histogram and enter the saved errors. Also check Display normal curve.<br>Calculate the Kolmogorov-Smirnov test (for $n \geq 50$) or Shapiro–Wilk test (for $n < 50$). ▶ Analyze ▶ Descriptive Statistics ▶ Explore ▶ Plots and check the **Normality plots with tests** box. |
| *Interpret the regression model* | |
| Consider the overall model fit. | Check the $R^2$ and adjusted $R^2$ and significance of the F-value. |
| Consider the effects of the independent variables separately. | Check the (standardized) $\beta$. Also check the sign of the $\beta$. 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. |

## 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 $\beta$ coefficients. Let us illustrate this by returning to our opening example. Imagine that we are trying to predict weekly supermarket sales (in USD) ($y$) 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:

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

**Table 7.3** 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 |
| | Price of product | −24003.037 | 16694.676 | −.241 | −1.438 | .162 |
| | Index of promotional activities | 44.227 | 13.567 | .547 | 3.260 | .003 |

a. Dependent Variable: Weekly sales in USD

If we estimate this model on a dataset, the estimated coefficients using regression analysis could be similar to those in Table 7.3.

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 USD and promotional activities at 50. Our expectation of the weekly sales would then be:

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

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 and worth the expenditure.

## Example

In the example, we take a closer look at the American Customer Satisfaction Index (*ACSI Data.sav*, Web Appendix → Chap. 7). Every year, the American Customer 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 end of 2007, the H.J. Heinz Company, Hershey Company, and Mars Incorporated 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.[5] 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 = yes, 0 = no).

The data included is a set of n = 1,640 responses provided by customers, but due to non-response, the actual number of responses for each variable is lower.

### Consider Data Requirements for Regression Analysis

First we need to check if our sample size is sufficient. By calculating descriptive statistics (▶ Analyze ▶ Descriptive Statistics ▶ 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

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

(1991). Although we have no further information on the sampling process, we assume that this has been correctly done. 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 we can only check this thoroughly after running a regression analysis. We will therefore discuss this later.

### 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 ▶ Analyze ▶ Regression ▶ 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 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

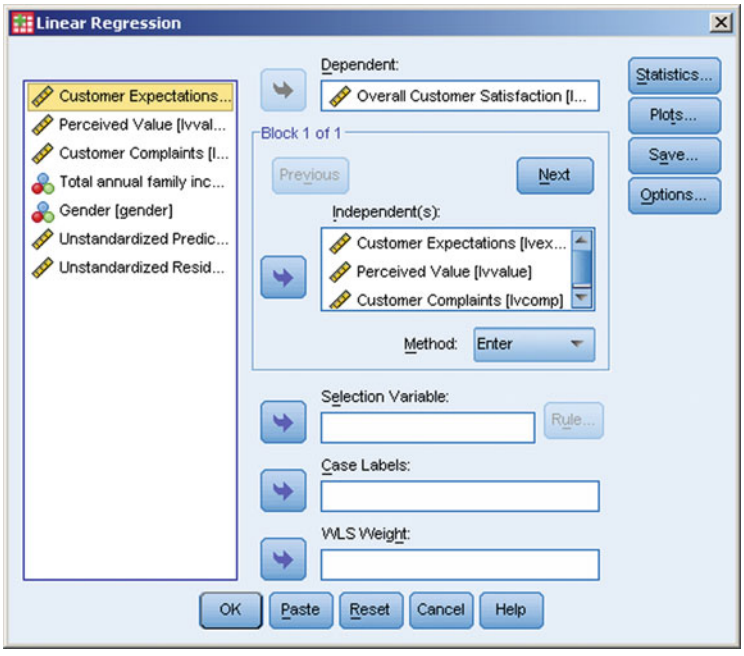


**Fig. 7.9** The Linear Regression dialog box

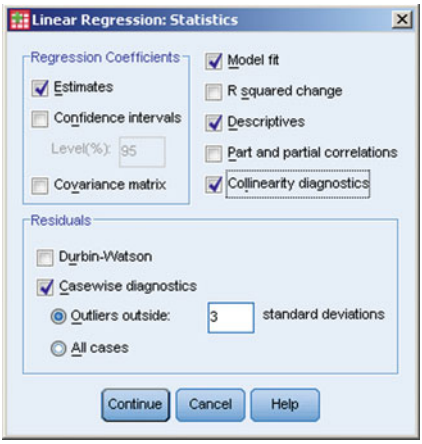


**Fig. 7.10** The statistics dialog box

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 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 **Continue**. 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.

### Test the Assumptions of Regression Analysis Using SPSS Regression Output

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

First, we check for linearity. If we create a scatter plot of *Overall Customer Satisfaction* against *Customer Expectations*, SPSS produces Fig. 7.12. This plot

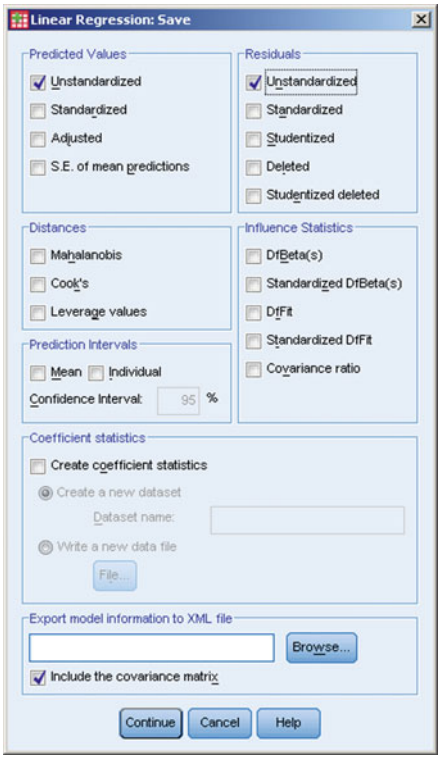


**Fig. 7.11** The Save options for regression analysis

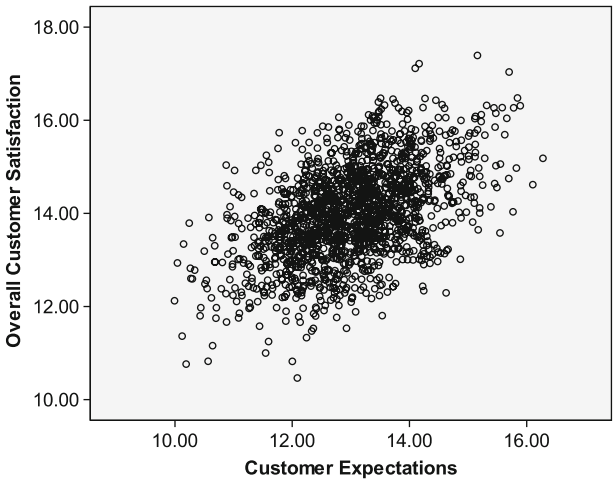


**Fig. 7.12** Overall customer satisfaction against customer expectations

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 few different values, such as *Customer Complaints* that only takes on the values of 0 and 1, linearity is hard to spot. Overall, the first assumption is fulfilled.

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 a 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 ▶ Graphs ▶ Legacy Dialogs ▶ Scatter/Dot (choose **Simple Scatter**). Enter the *Unstandardized Residual* to the **Y-axis** and put *Overall Customer Satisfaction* 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 we have set Casewise diagnostics to "Outliers outside: 2 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

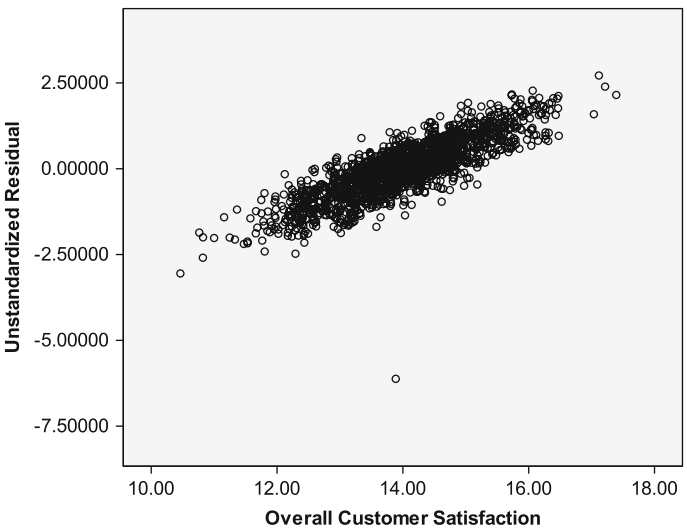


**Fig. 7.13** A plot of the errors against the predicted values

**Table 7.4** Casewise diagnostics[a]

**Casewise Diagnostics[a]**

| Case Number | Std. Residual | Weekly sales in USD | Predicted Value | Residual |
|---|---|---|---|---|
| 8 | 2.466 | 14134.00 | 8984.7644 | 5149.23563 |
| 18 | 2.241 | 10512.00 | 5832.2042 | 4679.79581 |

a. Dependent Variable: Weekly sales in USD

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? If we have many observations (as in this example) and only few outliers, we should avoid deleting these. However, case 257 seems to be very far away from the other observations (also see Fig. 7.13) and should be deleted. The other potential outliers appear to be simply part of the data, and don't much influence results, thus they should be included.

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.

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 ▶ Graphs ▶ Legacy Dialogs ▶ 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 7.14 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 going to ▶ 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.

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.

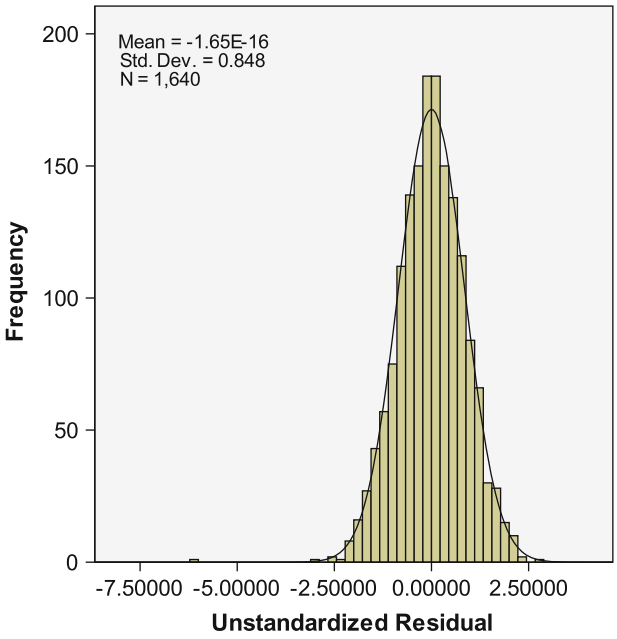


**Fig. 7.14** Histogram of the errors with a standard normal curve

**Table 7.5** Output produced by the Kolmogorov–Smirnov test

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Unstandardized Residual | .016 | 1639 | .200* | .998 | 1639 | .041 |

a. Lilliefors Significance Correction

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

## Interpret the Regression Model Using SPSS Regression Output

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.

**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 |

**Table 7.7** Correlation matrix

**Correlations**

| | | Overall Customer Satisfaction | Customer Expectations | Perceived Value | Customer Complaints |
|---|---|---|---|---|---|
| Pearson Correlation | Overall Customer Satisfaction | 1.000 | .492 | .766 | −.144 |
| | Customer Expectations | .492 | 1.000 | .478 | −.073 |
| | Perceived Value | .766 | .478 | 1.000 | −.137 |
| | Customer Complaints | −.144 | −.073 | −.137 | 1.000 |
| Sig. (1-tailed) | Overall Customer Satisfaction | . | .000 | .000 | .000 |
| | Customer Expectations | .000 | . | .000 | .001 |
| | Perceived Value | .000 | .000 | . | .000 |
| | Customer Complaints | .000 | .001 | .000 | . |
| N | Overall Customer Satisfaction | 1639 | 1639 | 1639 | 1639 |
| | Customer Expectations | 1639 | 1639 | 1639 | 1639 |
| | Perceived Value | 1639 | 1639 | 1639 | 1639 |
| | Customer Complaints | 1639 | 1639 | 1639 | 1639 |

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.** Furthermore, 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 $R^2$ provided in Table 7.9 seems highly 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 $R^2$ and adjusted $R^2$ are similar. If the adjusted $R^2$ 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

**Table 7.8** Variables used and regression method

**Variables Entered/Removed[b]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | Customer Complaints, Customer Expectations, Perceived Value[a] | . | Enter |

a. All requested variables entered.

b. Dependent Variable: Overall Customer Satisfaction

**Table 7.9** The model summary[a]

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .780[a] | .609 | .608 | .62756 |

a. Predictors: (Constant), Customer Complaints, Customer Expectations, Perceived Value

b. Dependent Variable: Overall Customer Satisfaction

**Table 7.10** ANOVA[a]

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 1001.261 | 3 | 333.754 | 847.453 | .000[a] |
| | 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

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | 5.124 | .215 | | 23.863 | .000 | | |
| | Customer Expectations | .164 | .018 | .163 | 9.257 | .000 | .771 | 1.296 |
| | Perceived Value | .677 | .018 | .683 | 38.484 | .000 | .761 | 1.314 |
| | Customer Complaints | –.092 | .037 | –.038 | –2.458 | .014 | .981 | 1.019 |

a. Dependent Variable: Overall Customer Satisfaction

the F-test. The result in Table 7.10 indicates that the regression model is significant (p-value $\leq 0.05$).

After assessing the overall model fit, it is time to look at the individual parameters. 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 three significant variables, those with p-values (under **Sig.**) 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 significant 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* (0.683) has clearly the strongest relationship with overall customer satisfaction. Therefore, this might be the first variable you want to increase if you aim to increase customer satisfaction. However, this assessment is not accurate since standardized $\beta s$ cannot be meaningfully compared when an independent variable is binary (as it is the case with *Customer Complaints*). Nevertheless, the comparison gives us a rough idea regarding the relative strengths of the independent variables' effects on the dependent variable.

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 standard errors are used to calculate the t-values. If we take the unstandardized coefficient 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 $\hat{y}$ for customers who complained compared to those who have not complained.

Overall, it seems that we have found a useful model that seems to satisfy the key assumptions of regression analysis.

### 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 ▶ Data ▶ 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%. 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 is 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.

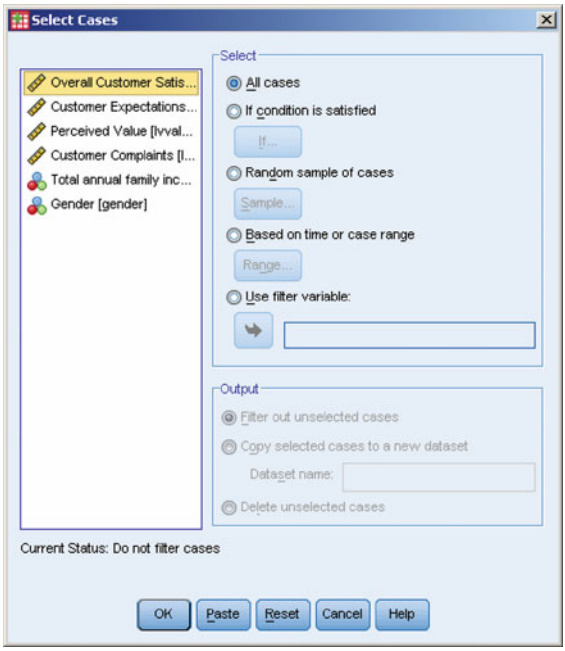


**Fig. 7.15** The select cases dialog box

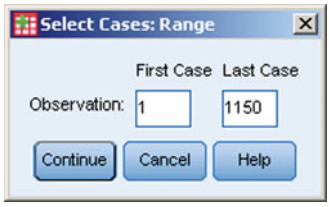


**Fig. 7.16** The select cases: range dialog box

## 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 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 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 Web Appendix (→ Chap. 7), we also include the original survey.[6]

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.

## Questions

1. Try to explain what regression analysis is in your own words.
2. Imagine you are asked to use regression analysis to explain the profitability of new supermarket products, such as the introduction of a new type of jam or yoghurt, in the first year of the launch. What independent variables would you use to explain the profitability of these new products?

[6]We would like to thank Dr. D.I. Gilliland and AgriPro for making the data and case available.

3. Imagine you are going to a client to present the findings of a regression model. The client believes that the regression model is a "black box" and that anything can be made significant. What would your reaction be?
4. I do not care about the assumptions – just give me the results! Please evaluate this statement. Do you agree?
5. Are all regression assumptions equally important? Please discuss assumptions
6. Using standardized $\beta s$, we can compare effects between different variables. Can we really compare apples and oranges after all? Please discuss.
7. Run the ACSI example without deleting the outlier observation (i.e. using the full dataset with 1,640 observations) and compare the results with those presented above. Explain why deviations occur.

## Further Readings

American Customer Satisfaction index at http://www.theacsi.org

*This website contains scores of the American Satisfaction Index.*

Garson's Statnotes page at http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

*This website provides an excellent overview on regression analysis from a technical perspective.*

Hair JF, Black WC, Babin BJ, Anderson RE (2010) Multivariate data analysis. A global perspective, 7th edn. Pearson Prentice Hall, Upper Saddle River, NJ

*This is an excellent book, which discusses many statistical terms from a theoretical perspective in a highly accessible manner.*

Nielsen at http://www.nielsen.com

*This is the website for Nielsen, one of the world's biggest market research companies. They publish many reports that use regression analysis.*

The Food Marketing Institute at http://www.fmi.org

*This website contains data, some of which can be used for regression analysis.*

## References

Cohen J (1994) The Earth is round (P <.05). Am Psychol 49(912):997–1003

Field A (2009) Discovering statistics using SPSS, 3rd edn. Sage, London

Green SB (1991) How many subjects does it take to do a regression analysis? Multivariate Behav Res 26:499–510

Greene WH (2007) Econometric analysis, 6th edn. Prentice Hall, Upper Saddle River, NJ

Hill C, Griffiths W, Lim GC (2008) Principles of econometrics, 3rd edn. Wiley, Hoboken, NJ

Kelley K, Maxwell SE (2003) Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. Psychol Methods 8(3):305–321

Ringle CM, Sarstedt M, Mooi EA (2010) Response-based segmentation using FIMIX-PLS. Theoretical foundations and an application to ACSI data. Ann Inf Syst 8:19–49