Regression Analysis

7.1 Introduction – 210

7.2 Understanding Regression Analysis – 211

7.3 Conducting a Regression Analysis – 214

7.3.1 Check the Regression Analysis Data Requirements – 214

7.3.2 Specify and Estimate the Regression Model – 217

7.3.3 Test the Regression Analysis Assumptions – 221

7.3.4 Interpret the Regression Results – 227

7.3.5 Validate the Regression Results – 232

7.3.6 Use the Regression Model – 233

7.4 Example – 238

7.5 Farming with AgriPro (Case Study) – 252

7.6 Review Questions – 254

References – 255

Electronic supplementary material
The online version of this chapter (https://doi.org/10.1007/978-3-662-56707-4_7) contains additional material that is available to authorized users. You can also download the "Springer Nature More Media App" from the iOS or Android App Store to stream the videos and scan the image containing the "Play button".

© Springer-Verlag GmbH Germany, part of Springer Nature 2019

M. Sarstedt, E. Mooi, *A Concise Guide to Market Research*, Springer Texts in Business
and Economics, https://doi.org/10.1007/978-3-662-56707-4_7

Learning Objectives

After reading this chapter you should understand:

- The basic concept of regression analysis.
- How regression analysis works.
- The requirements and assumptions of regression analysis.
- How to specify a regression analysis model.
- How to interpret regression analysis results.
- How to predict and validate regression analysis results.
- How to conduct regression analysis with SPSS.
- How to interpret regression analysis output produced by SPSS.

Keywords

Adjusted R^2 • Akaike information criterion • Autocorrelation • Bayes information criterion • Binary logistic regression • Bootstrapping • Coefficient of determination • Collinearity • Constant • Control variables • Cross-validation • Dependent variables • Dummy variables • Durbin-Watson test • Error • Error sum of squares • Estimation sample • *F*-test • Heteroscedasticity • Homoscedasticity • Independent variables • Intercept • Moderation analysis • Multicollinearity • Multinomial logistic regression • Multiple regression • Nested models • Ordinary least squares • Outliers • Ramsey's RESET test • Regression sum of squares • Residual • R^2 • Simple regression • Split-sample validation • Standard error • Standardized effects • Tolerance • Unstandardized effects • Validation Sample • Variance inflation factor • Weighted Least Squares

7.1 Introduction

Regression analysis is one of the most frequently used analysis techniques in market research. It allows market researchers to analyze the relationships between **dependent variables** and **independent variables**. In marketing applications, the dependent variable is the outcome we care about (e.g., sales), while we use the independent variables to achieve those outcomes (e.g., pricing or advertising). The key benefits of using regression analysis are it allows us to:

1. Calculate if one independent variable or a set of independent variables has a significant relationship with a dependent variable.
2. Estimate the relative strength of different independent variables' effects on a dependent variable.
3. Make predictions.

Knowing what the effects of independent variables on dependent variables are, helps market researchers in many different ways. For example, this knowledge can help guide spending if we know promotional activities relate strongly to sales.

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

Regression analysis can also help us make predictions. For example, if we have estimated a regression model by using data on the weekly supermarket sales of a brand of milk in dollars, the milk price (which changes with the season and supply), as well as an index of promotional activities (comprising product placement, advertising, and coupons), the results of the regression analysis could answer the question: what would happen to the sales if the prices were to increase by 5 % and the promotional activities by 10 %? Such answers help (marketing) managers make sound decisions. Furthermore, by calculating various scenarios, such as price increases of 5, 10, and 15 %, managers can evaluate marketing plans and create marketing strategies.

7.2 Understanding Regression Analysis

In the previous paragraph, we briefly discussed what regression can do and why it is a useful market research tool. We now provide a more detailed discussion. Look at ◘ Fig. 7.1: It plots a dependent (y) variable (the weekly sales of a brand of milk in dollars against an independent (x_1) variable (an index of promotional activities). Regression analysis is a way of fitting a "best" line through a series of observations. With a "best" line we mean one that is fitted in such a way that it minimizes the sum of the squared differences between the observations and the line itself. It is important to know that the best line fitted by means of regression analysis is not necessarily the true line (i.e., the line that represents the population). Specifically, if we have data issues, or fail to meet the regression assumptions (discussed later), the estimated line may be biased.

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

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

What does this mean? The y represents the dependent variable, which is the outcome you are trying to explain. In ◘ Fig. 7.1, we plot the dependent variable on the vertical axis. The α represents the **constant** (or **intercept**) of the regression model, and indicates what your dependent variable would be if the independent variable were zero. In ◘ Fig. 7.1, you can see the constant is the value where the fitted straight (sloping) line crosses the y-axis, which is at 2463.963. Thus, if the index of promotional activities is zero, we expect the weekly supermarket sales of a specific milk brand to be $2464. It may not always be realistic to assume that independent variables are zero (prices are, after all, rarely zero), but the constant should always be included to ensure the regression model's best possible fit with the data.

The independent variable is indicated by x_1, while the β_1 (pronounced beta) indicates its (regression) coefficient. This coefficient represents the slope of the line, or the slope of the diagonal grey line in ◘ Fig. 7.1. A positive β_1 coefficient indicates an upward sloping regression line, while a negative β_1 coefficient indicates a downward sloping line. In our example, the line slopes upward. This makes sense, since sales tend to increase with an increase in promotional activities. In our example, we estimate the β_1 as 54.591, meaning

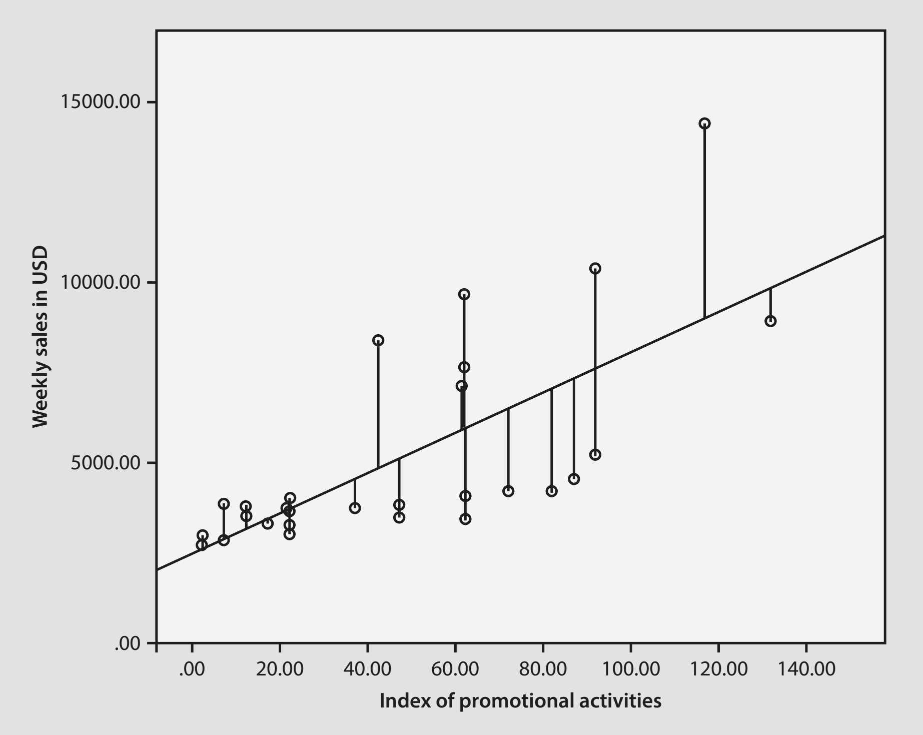


Fig. 7.1 A visual explanation of regression analysis

that if we increase the promotional activities by one unit, the weekly supermarket sales of a brand of milk will go up by an average of $54.591. This ß$_1$ value has a degree of associated uncertainty called the **standard error**. This standard error is assumed to be normally distributed. Using a *t*-test (see ▶ Chap. 6), we can test if the ß$_1$ is indeed significantly different from zero.

The last element of the notation, the *e*, denotes the equation **error** (also called the **residual** or **disturbance term**). The error is the distance between each observation and the best fitting line. To clarify what a regression error is, examine Fig. 7.1 again. The error is the difference between the regression line (which represents our regression prediction) and the actual observation (indicated by each dot). The predictions made by the "best" regression line are indicated by $\hat{y}$ (pronounced y-hat). Thus, the error of each observation is:[1]

$$e = y - \hat{y}$$

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

In the example above, we have only one independent variable. We call this **simple regression**. If we include multiple independent variables, we call this **multiple regression**. The notation for multiple regression is like that of simple regression. If we were to have two independent variables, for example the price (x_1), and an index of promotional activities (x_2), our notation would be:

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

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

The Explained Visually webpage offers an excellent visualization of how regression analysis works.

© Kurt Kleemann/stock.adobe.com
http://setosa.io/ev/ordinary-least-squares-regression/

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

[2] This only applies to the standardized βs.

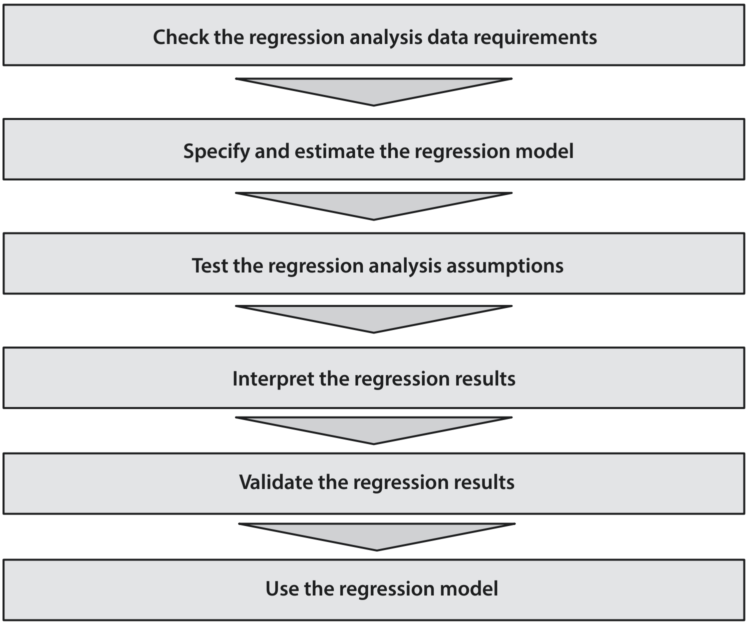

Check the regression analysis data requirements
Specify and estimate the regression model
Test the regression analysis assumptions
Interpret the regression results
Validate the regression results
Use the regression model

Fig. 7.2 Steps involved in a regression analysis

7.3 Conducting a Regression Analysis

7.3.1 Check the Regression Analysis Data Requirements

Various data requirements must be taken into consideration before we undertake a regression analysis. These include the:

- Sample size,
- variables need to vary,
- scale type of the dependent variable, and
- collinearity.

We discuss each requirement in turn.

7.3.1.1 Sample Size

The first data requirement is that we need an "acceptable" sample size. "Acceptable" relates to a sample size that gives you a good chance of finding significant results if they are possible (i.e., the analysis achieves a high degree of statistical power; see ► Chap. 6). There are two ways to calculate "acceptable" sample sizes.

- The first, formal, approach is a power analysis. As mentioned in ► Chap. 6 (Box 6.2), these calculations require you to specify several parameters, such as the expected effect size and the maximum type I error you want to allow for. Generally, you should set the power to 0.80, which is an acceptable level. A power level of 0.80 means there is an 80 % probability of deciding that an effect will be significant, if it is indeed significant. Kelley and Maxwell (2003) discuss sample size requirements in far more detail. G*power—a free program available at http://www.gpower.hhu.de —is commonly used to calculate sample sizes precisely. SPSS also provides an add-on module called "Sample Power," which can be used to carry out such analyses.

- The second approach is by using rules of thumb. These rules are not specific or precise, but are easy to apply. Green (1991); VanVoorhis and Morgan (2007) suggest that if you want to test for individual parameters' effect (i.e., whether one coefficient is significant or not), you need a sample size of 104 + *k*. Thus, if you have ten independent variables, you need 104 + 10 = 114 observations. Note that this rule of thumb is best applied when you have a small number of independent variables, less than 10 and certainly less than 15. VanVoorhis and Morgan (2007) add that having at least 30 observations per variable (i.e. 30 *k*) allows for detecting smaller effects (an expected R^2 of 0.10 or smaller) better.

7.3.1.2 Variables Need to Vary

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

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

7.3.1.3 Scale Type of the Dependent Variable

The third data requirement is that the dependent variable needs to be interval or ratio scaled (► Chap. 3 discusses scaling). If the data are not interval or ratio scaled, alternative types of regression should be used. You should use **binary logistic regression** if the dependent variable is binary and only takes two values (zero and one). If the dependent variable is a nominal variable with more than two levels, you should use **multinomial logistic regression**. This should, for example, be used if you want to explain why people prefer product A over B or C. We do not discuss these different methods in this chapter, but they are related to regression. For a discussion of regression methods for dependent variables measured on a nominal or ordinal scale, see Field (2013).

7.3.1.4 Collinearity

The last data requirement is that no or little collinearity should be present.[3] **Collinearity** is a data issue that arises if two independent variables are highly correlated. Perfect collinearity occurs if we enter two or more independent variables containing exactly the same

[3] This is only a requirement if you are interested in the regression coefficients, which is the dominant use of regression. If you are only interested in prediction, collinearity is not important.

information, therefore yielding a correlation of 1 or −1 (i.e., they are perfectly correlated). Perfect collinearity may occur if you enter the same independent variable twice, or if one variable is a linear combination of another (e.g., one variable is a multiple of another variable, such as sales in units and sales in thousands of units). If this occurs, regression analysis cannot estimate one of the two coefficients. In practice, however, weaker forms of collinearity are common. For example, if we study what drives supermarket sales, variables such as price reductions and promotions are often used together. If this occurs very often, the variables price and promotion may be collinear, which means there is little uniqueness or new information in each of the variables. The problem with having collinearity is that it tends to regard significant parameters as insignificant. Substantial collinearity can even lead to sign changes in the regression coefficients' estimates. When three or more variables are strongly related to each other, we call this **multicollinearity**.

Fortunately, collinearity is relatively easy to detect by calculating the **variance inflation factor** (**VIF**).[4] The VIF indicates the effect on the standard error of the regression coefficient for each independent variable. Specifically, the square root of the VIF indicates you how much larger the standard error is, compared to if that variable were uncorrelated with all other independent variables in the regression model. Generally, a VIF of 10 or above indicates that (multi) collinearity is a problem (Hair et al. 2019).[5] Some research suggests that VIFs far above 10—such as 20 or 40—can be acceptable if the sample size is large and the R^2 (discussed later) is high (0.90 or more) (O'brien 2007). Conversely, if the sample sizes are below 200 and the R^2 is low (0.25 or less), collinearity is more problematic (Mason and Perreault 1991). Consequently, in such situations, lower VIF values—such as 5—should be the maximum.

You can remedy collinearity in several ways but each of these have costs. If perfect collinearity occurs, drop one of the perfectly overlapping variables. If weaker forms of collinearity occur, you can utilize three approaches to reduce collinearity (O'brien 2007):

- The first option is to use principal component or factor analysis on the collinear variables (see ► Chap. 8). By using principal component or factor analysis, you create a small number of factors that comprise most of the original variables' information, but are uncorrelated. If you use factors, collinearity no longer an issue.
- The second option is to re-specify the regression model by removing highly correlated variables. Which variables should you remove? If you create a correlation matrix of all the independent variables entered in the regression model, you should first focus on the variables that are most strongly correlated (see ► Chap. 5 for how to create a correlation matrix). First try removing one of the two most strongly correlated variables. The one you should remove depends on your research problem—retain the most relevant variable of the two.

[4] A related measure is the **tolerance**, which is 1/VIF and calculated as $1/(1-R^2)$.

[5] The VIF is calculated using a completely separate regression analysis. In this regression analysis, the variable for which the VIF is calculated is regarded as a dependent variable and all other independent variables are regarded as independents. The R^2 that this model provides is deducted from 1 and the reciprocal value of this sum (i.e., $1/(1-R^2)$) is the VIF. The VIF is therefore an indication of how much the regression model explains one independent variable. If the other variables explain much of the variance (the VIF is larger than 10), collinearity is likely a problem.

- The third option is not to do anything. In many cases removing collinear variables does not reduce the VIF values significantly. Even if we do, we run the risk of mis-specifying the regression model (see Box 7.1 for details). Given the trouble researchers go through to collect data and specify a regression model, it is often better to accept collinearity in all but the most extreme cases. In this case, however, we should be cautious with the interpretation of the regression coefficients because their values and significance levels may be biased.

7.3.2 Specify and Estimate the Regression Model

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

7.3.2.1 Model Specification

The model specification step involves choosing the variables to use. The regression model should be simple yet complete. To quote Albert Einstein: “Everything should be made as simple as possible, but not simpler!” How do we achieve this? By focusing on our ideas of what relates to the dependent variable of interest, the availability of data, client requirements, and prior regression models. For example, typical independent variables that explain the sales of a product include the price and promotions. When available, in-store advertising, competitors’ prices, and promotions are usually also included. Market researchers may, of course, choose different independent variables for other applications. Omitting important variables (see Box 7.1) has substantial implications for the regression model, so it is best to be inclusive. A few practical suggestions:

- If you have many variables available in the data that overlap in terms of how they are defined— such as satisfaction with the waiter/waitress and with the speed of service—try to pick the variable that is most distinct or relevant for the client. Alternatively, you could conduct a principal component or factor analysis (see ► Chap. 8) first and use the factors as the regression analysis’s independent variables.

Box 7.1 Omitting relevant variables

Omitting key variables from a regression model can lead to biased results. Imagine that we want to explain weekly sales by only referring to promotions. From the introduction, we know the *ß* of the regression model only containing promotions is estimated as 54.591. If we add the variable price (arguably a key variable), the estimated *ß* of promotions drops to 42.266. As can be seen, the difference between the estimated *ßs* in the two models (i.e., with and without price) is 12.325, suggesting that the “true” relationship between promotions and sales is weaker than in a model with only one independent variable. This example shows that omitting important independent variables leads to biases in the value of the estimated *ßs*. That is, if we omit a relevant variable x_2 from a regression model that only includes x_1, we cause a bias in the $ß_1$ estimate. More precisely, the $ß_1$ is likely to be inflated, which means that the estimated value is higher than it should be. Thus, the $ß_1$ itself is biased because we omit x_2!

- If you expect to need a regression model for different circumstances, you should make sure that the independent variables are the same, which will allow you to compare the models. For example, temperature can drive the sales of some supermarket products (e.g., ice cream). In some countries, such as Singapore, the temperature is relatively constant, so including this variable is not important. In other countries, such as Germany, the temperature can fluctuate far more. If you are intent on comparing the ice cream sales in different countries, it is best to include variables that may be relevant to all the countries you want to compare (e.g., by including temperature, even if it is not very important in Singapore).
- Consider the type of advice you want to provide. If you want to make concrete recommendations regarding how to use point-of-sales promotions and free product giveaways to boost supermarket sales, both variables need to be included.
- Take the sample size rules of thumb into account. If practical issues limit the sample size to below the threshold that the rules of thumb recommend, use as few independent variables as possible. Larger sample sizes allow you more freedom to add independent variables, although they still need to be relevant.
- If you have ordinal variables, you need to construct **dummy variables**. Dummy variables indicate categories or subgroups using a 0 (absent) or 1 (present) value. Dummies can be used to understand if for example ice cream sales differ between three countries, France, Germany, and Singapore. We need only two variables to indicate the groups since if sales observations do not come from Germany (=0) or Singapore (=0), they need to be from France. Thus, we have used France as the base category. Sometimes the choice of base category is not so important but if we have a base category where a characteristic is absent (e.g., no advertising compared to three dummies that indicate three different forms of advertising) it is useful to set this as the base category (by omitting a dummy for this category). Note that we always construct one dummy fewer than the number of groups and that it is only a decision for which category to omit the dummy! We can enter dummy variables to a regression model to estimate if there are differences between categories or subgroups (such as countries or types of advertising).

7.3.2.2 Model Estimation

Model estimation refers to how we estimate a regression model. The most common method of estimating regression models is **ordinary least squares** (**OLS**). OLS fits a regression line to the data that minimizes the sum of the squared distances to it. These distances are squared to stop negative distances (i.e., below the regression line) from cancelling out positive distances (i.e., above the regression line), because squared values are always positive. Moreover, by using the square, we emphasize observations that are far from the regression much more, while observations close to the regression line carry very little weight. The rule to use squared distances is an effective (but also arbitrary) way of calculating the best fit between a set of observations and a regression line (Hill et al. 2011). If we return to ◘ Fig. 7.1, we see the vertical "spikes" from each observation to the regression line. OLS estimation is aimed at minimizing the squares of these spikes.

We use the data behind ◘ Fig. 7.1—as shown in ◘ Table 7.1—to illustrate the method with which OLS regressions are calculated. This data has 30 observations, with information

Table 7.1 Regression data

Week	Sales	Price	Promotion
1	3454	1.10	12.04
2	3966	1.08	22.04
3	2952	1.08	22.04
4	3576	1.08	22.04
5	3692	1.08	21.42
6	3226	1.08	22.04
7	3776	1.09	47.04
8	14,134	1.05	117.04
9	5114	1.10	92.04
10	4022	1.08	62.04
11	4492	1.12	87.04
12	10,186	1.02	92.04
13	7010	1.08	61.42
14	4162	1.06	72.04
15	3446	1.13	47.04
16	3690	1.05	37.04
17	3742	1.10	12.04
18	7512	1.08	62.04
19	9476	1.08	62.04
20	3178	1.08	22.04
21	2920	1.12	2.04
22	8212	1.04	42.04
23	3272	1.09	17.04
24	2808	1.11	7.04
25	2648	1.12	2.04
26	3786	1.11	7.04
27	2908	1.12	2.04
28	3395	1.08	62.04
29	4106	1.04	82.04
30	8754	1.02	132.04

on the supermarket's sales of a brand of milk (*sales*), the price (*price*), and an index of promotional activities (*promotion*) for weeks 1–30. This dataset is small and only used to illustrate how OLS estimates are calculated. The data can be downloaded (↓ Web Appendix → Downloads), but are also included in ◘ Table 7.1.

To estimate the effect of *price* and *promotion* on *sales*, we need to calculate the *ß*s, of which the estimate is noted as $\hat{\beta}$ (pronounced as beta-hat). The $\hat{\beta}$ indicates the estimated association between each independent variable (*price* and *promotion*) and the dependent variable *sales*. We can estimate $\hat{\beta}$ as follows:

$$\hat{\beta} = (x^T x)^{-1}\, x^T y$$

In this equation, to solve the $\hat{\beta}$ s, we first multiply the transposed matrix indicated as x^T. This matrix has three elements, a vector of 1s, which are added to estimate the intercept and two vectors of the independent variables *price* and *promotion*. Together, these form a 30 × 3 matrix. Next, we multiply this matrix with the untransposed matrix, indicated as *x*, consisting of the same elements (as a 3 × 30 matrix). This multiplication results in a 3 × 3 matrix of which we calculate the inverse, indicated by the power of −1 in the equation. This also results in a 3 × 3 matrix $(x^T x)^{-1}$. Next, we calculate $x^T y$, which consists of the 30 × 3 matrix and the vector with the dependent variables' observations (a 1 × 30 matrix). In applied form:

$$x = \begin{bmatrix} 1 & 1.10 & 12.04 \\ 1 & 1.08 & 22.04 \\ \vdots & \vdots & \vdots \\ 1 & 1.02 & 132.04 \end{bmatrix},$$

$$x^T = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1.10 & 1.08 & \ldots & 1.02 \\ 12.04 & 22.04 & \ldots & 132.04 \end{bmatrix},$$

$$(x^T x)^{-1} = \begin{bmatrix} 77.97 & -70.52 & -0.04 \\ -70.52 & 63.86 & 0.03 \\ -0.04 & 0.03 & 0.00 \end{bmatrix},^{6}$$

$$x^T y = \begin{bmatrix} 147615.00 \\ 158382.64 \\ 8669899.36 \end{bmatrix},$$

[6] This term can be calculated manually, but also by using the function *mmult* in Microsoft Excel where $x^T x$ is calculated. Once this matrix has been calculated, you can use the *minverse* function to arrive at $(x^T x)^{-1}$.

$$\text{Hence,}\quad (x^T x)^{-1} \cdot x^T y = \begin{bmatrix} 30304.05 \\ -25209.86 \\ 42.27 \end{bmatrix}.$$

This last matrix indicates the estimated βs (i.e., $\hat{\beta}$) with 30,304.05 representing the intercept, −25,209.86 representing the effect of a one-unit increase in the *price* on *sales*, and 42.27 the effect of a one-unit increase in *promotions* on *sales*.

As discussed before, each $\hat{\beta}$ also has a standard error. This standard error can be expressed in standard deviations and, as seen in the discussion in ► Chap. 5, values outside the $\pm 1.96 \cdot t$-value from the $\hat{\beta}$ indicate the significance of two-tailed tests. If this range includes a value of 0, the $\hat{\beta}$ is said to be *insignificant*. If it excludes 0, the $\hat{\beta}$ is said to be *significant*.

While OLS is an effective estimator, there are alternatives that work better in specific situations. These situations occur if we violate one of the regression assumptions. For example, if the regression errors are heteroscedastic (discussed in ► Sect. 7.3.3.3), we need to account for this by, for example, using **Weighted Least Squares** (*WLS*). We briefly discuss when WLS should be used in this chapter. If the expected mean error of the regression model is not zero, estimators such as two-staged least squares (2SLS) can be used in specific situations. There are many more estimators, but these are beyond the scope of this book. Greene (2011) discusses these and other estimation procedures in detail.

7.3.3 Test the Regression Analysis Assumptions

We have already discussed several issues that determine whether running a regression analysis is useful. We now discuss regression analysis assumptions. If a regression analysis fails to meet its assumptions, it can provide invalid results. Four regression analysis assumptions are required to provide valid results:
1. The regression model can be expressed linearly,
2. The regression model's expected mean error is zero,
3. The errors' variance is constant (homoscedasticity), and
4. The errors are independent (no autocorrelation).

There is a fifth assumption which is optional. If we meet this assumption, we have information on how the regression parameters are distributed, which allows straightforward conclusions regarding their significance. If the regression analysis fails to meet this assumption, the regression model will still be accurate, but we cannot rely accurately on the standard errors (and *t*-values) to determine the regression parameters' significance.

5. The errors need to be approximately normally distributed.

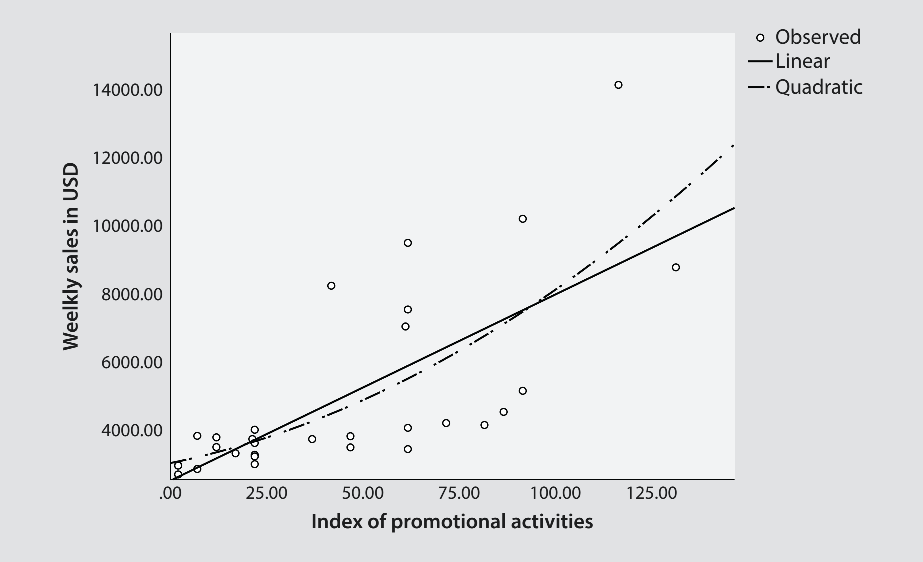
We next discuss these assumptions and how we can test each of them.

7.3.3.1 First Assumption: Linearity

The first assumption means that we can write the regression model as $y = \alpha + \beta_1 x_1 + e$. Thus, non-linear relationships, such as $\beta_1^2 x_1$, are not permissible. However, logarithmic expressions, such as log(x_1), are possible as the regression model is still specified linearly. If you can write a model whose regression parameters (the *ß*s) are linear, you satisfy this assumption.

A separate issue is whether the relationship between the independent variable *x* and the dependent variable *y* is linear. You can check the linearity between *x* and *y* variables by plotting the independent variables against the dependent variable. Using a scatter plot, we can then assess whether there is some type of non-linear pattern. ■ Figure 7.3 shows such a plot. The straight, sloping line indicates a linear relationship between *sales* and *promotions*. For illustration purposes, we have also added a curved upward sloping line. This dashed line corresponds to a x_1^2 transformation. It visually seems that the quadratic line fits the data best. If we fail to identify non-linear relationships as such, our regression line does not fit the data well, as evidenced in a low model fit (e.g., the R^2, which we discuss later) and nonsignificant effects. After transforming x_1 by squaring it (or using any other transformation), you still satisfy the assumption of specifying the regression model linearly, despite the non-linear relationship between *x* and *y*. We discuss details of transformations in Box 7.3 later in this chapter.

Ramsey's RESET test is a specific linearity test (Ramsey 1969; Cook and Weisberg 1983). This test includes the squared values of the independent variables (i.e., x_1^2) and third powers (i.e., x_1^3), and tests if these are significant (Baum 2006).[7] While this test can detect these specific types of non-linearities, it does not indicate which variable or variables have



■ **Fig. 7.3** Different relationships between promotional activities and weekly sales

[7] The test also includes the predicted values squared and to the power of three.

a non-linear relationship with the dependent variable. Sometimes this test is (falsely) called a test for omitted variables, but it actually tests for non-linearities.

7.3.3.2 Second Assumption: Expected Mean Error Is Zero

The second assumption is that the expected (not the estimated!) mean error is zero. If we do not expect the sum of the errors to be zero, we obtain a biased line. That is, we have a line that consistently overestimates or underestimates the true relationship. This assumption is not testable by means of statistics, as OLS always renders a best line with a calculated mean error of exactly zero. This assumption is important, because if the error's expected value is not zero, there is additional information in the data that has not been used in the regression model. For example, omitting important variables, as discussed in Box 7.1, or autocorrelation may cause the expected error to no longer be zero (see ► Sect. 7.3.3.4).

7.3.3.3 Third Assumption: Homoscedasticity

The third assumption is that the errors' variance is constant, a situation we call **homoscedasticity**. Imagine that we want to explain various supermarkets' weekly sales in dollars. Large stores obviously have a far larger sales spread than small supermarkets. For example, if you have average weekly sales of $50,000, you might see a sudden jump to $60,000, or a fall to $40,000. However, a very large supermarket could see sales move from an average of $5 million to $7 million. This causes the weekly sales' error variance of large supermarkets to be much larger than that of small supermarkets. We call this non-constant variance **heteroscedasticity**. If we estimate regression models on data in which the variance is not constant, they will still result in correct βs. However, the associated standard errors are likely to be too large and may cause some βs to not be significant, although they actually are.

◘ Figure 7.4 provides a visualization of heteroscedasticity. As the dependent variable increases, the error variance also increases. A simple scatterplot (see ► Chap. 5) with the dependent variable on the y-axis and the error plotted on the x-axis can visualize heteroscedasticity. If heteroscedasticity is an issue, the points are typically funnel shaped, displaying more (or less) variance as the independent variable increases (decreases). Note that the variance can also first go up and then down (diamond shaped) or down and then up (diabolo shaped). These funnel shapes are typical of heteroscedasticity and indicate that the error variance changes as a function of the dependent variable. If these plots reveal heteroscedasticity as an issue, we can deal with it in two different ways described next.

First, WLS may be used. Compared to OLS, where each observation has an equal weight (see ► Sect. 7.3.2.2), WLS "weights" the regression line such that observations with a smaller variance are given greater weight in determining the regression coefficients. The difficult part is to understand which variable drives the heteroscedasticity. Returning to the store sales example above, if a variable store size were available, it would address the issue that weekly sales' error variance of large supermarkets is much larger than that of small supermarkets. This would make it a useful weighting variable. Only use WLS if there is a clear indication of heteroscedasticity and if you have a good weighting variable. Before using

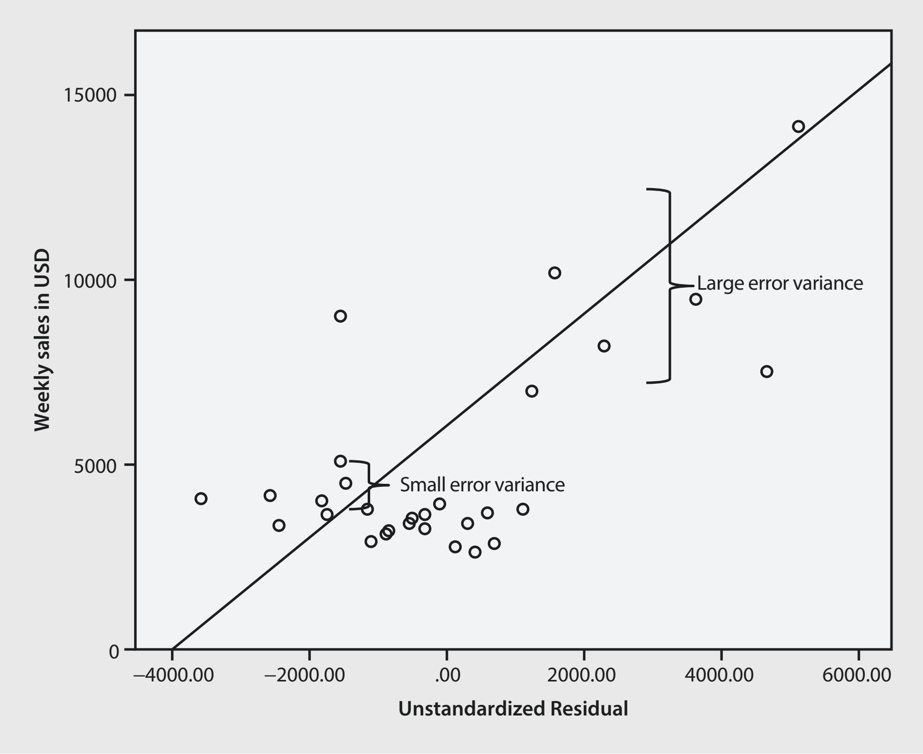


Fig. 7.4 An example of heteroscedasticity

WLS, try adding the weight variable to the original regression model to see if heteroscedasticity is still a problem. Often heteroscedasticity arises because of omitted variables.

Second, **bootstrapping** may be used. If we expect that our regression model suffers from heteroscedasticity, bootstrapping is an alternative. Bootstrapping takes different samples from the dataset and estimates many regression models (typically 1000) from this dataset. For each sample, the regression model is calculated and these different values are weighted into bootstrapped $\hat{\beta}$ s and standard errors. These bootstrapped $\hat{\beta}$ s and standard errors give an indication of how robust the results are and if the interpretation of the OLS and bootstrapped $\hat{\beta}$ s and standard errors are the same, heteroscedasticity, is unlikely to be a major concern. Moreover, the bootstrapped $\hat{\beta}$ s and standard errors give an indication of the true values of both when heteroscedasticity is controlled for.

7.3.3.4 Fourth Assumption: No Autocorrelation

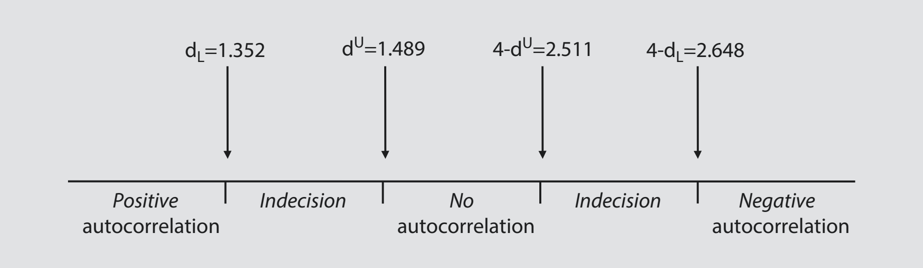
The fourth assumption is that the regression model errors are independent; that is, the error terms are uncorrelated for any two observations. Imagine that you want to explain the supermarket sales of a brand of milk by using the previous week's sales of that milk. It is very likely that if sales increased last week, they will also increase this week. This may be due to, for example, the growing economy, an increasing appetite for milk, or other reasons that underlie the growth in supermarket sales of milk. This issue is called **autocorrelation** and means that regression errors are correlated positively (or negatively) over time. For example, the data in ◘ Table 7.1 are taken from weeks 1 to 30, which means they have a time component.

We can identify the presence of autocorrelation by using the **Durbin-Watson (D-W) test** (Durbin and Watson 1951). The D-W test assesses whether there is autocorrelation by testing the null hypothesis of no autocorrelation, which is tested for negative autocorrelation against a lower and upper bound and for positive autocorrelation against a lower and upper bound. If we reject the null hypothesis of no autocorrelation, we find support for an alternative hypothesis that there is some degree of positive or negative autocorrelation. Essentially, there are four situations, which we indicate in ◘ Fig. 7.5.

1. The errors may be positively related (called positive autocorrelation). This means that if we have observations over time, we observe that positive errors are generally followed by positive errors and negative errors by negative errors. For example, supermarket sales usually increase over certain time periods (e.g., before Christmas) and decrease during other periods (e.g., the summer holidays).
2. If positive errors are commonly followed by negative errors and negative errors by positive errors, 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 to simply go home on time.
3. If no systematic pattern of errors occurs, we have no autocorrelation. This absence of autocorrelation is required to estimate standard (OLS) regression models.
4. The D-W values may fall between the lower and upper critical value. If this occur, the test is inconclusive.

The situation that occurs depends on the interplay between the D-W test statistic (d) and the lower (d_L) and upper (d^U) critical value:

- If the test statistic is lower than the lower critical value ($d < d_L$), we have positive autocorrelation.
- If the test statistic is higher than 4 minus the lower critical value ($d > 4-d_L$), we have negative autocorrelation.
- If the test statistic falls between the upper critical value and 4 minus the upper critical value ($d^U < d < 4-d^U$), we have no autocorrelation.



◘ **Fig. 7.5** Durbin-Watson test values ($n = 30$, $k = 1$)

- If the test statistic falls between the lower and upper critical value ($d_L < d < d^U$), or it falls between 4 minus the upper critical value and 4 minus the lower critical value ($4–d^U < d < 4–d^L$), the test does not inform on the presence of autocorrelation and is undecided.

The critical values d_L and d^U can be found on the website accompanying this book (↓ Web Appendix → Downloads). From this table, you can see that the lower critical value d_L of a model with one independent variable and 30 observations is 1.352 and the upper critical value d^U is 1.489. ◘ Figure 7.5 shows the resulting intervals. Should the D-W test indicate autocorrelation, you should use models that account for this problem, such as panel and time series models. We do not discuss these methods in this book, but Hill et al. (2011) is a useful source of further information.

© echoevg/Getty Images/iStock
https://www.guide-marketresearch.com/app/download/13488671527/SPSS+3rd_Chapter+7_DW+Test.pdf?t=1516713141

7.3.3.5 Fifth (Optional) Assumption: Error Distribution

The fifth, optional, assumption is that the regression model errors are approximately normally distributed. If this is not the case, the *t*-values may be incorrect. However, even if the regression model errors are not normally distributed, the regression model still provides good estimates of the coefficients. Consequently, we consider this assumption an optional one. Potential reasons for regression errors being non-normally distributed include **outliers** (discussed in ► Chap. 5) and a non-linear relationship between the dependent and one or more independent variable(s) as discussed in ► Sect. 7.3.3.1.

There are two main ways of checking for normally distributed errors: you can use plots or carry out a formal test. Formal tests of normality include the Shapiro-Wilk tests (see ► Chap. 6), which needs to be run on the saved errors. A formal test may indicate non-normality and provide absolute standards. However, formal test results reveal little about the source of non-normality. A histogram with a normality plot may can help assess why errors are non-normally distributed (see ► Chap. 5 for details). Such plots are easily explained and interpreted and may suggest the source of non-normality (if present).

7.3.4 Interpret the Regression Results

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

7.3.4.1 Overall Model Fit

Assessing the overall model fit starts by interpreting the ***F*-test**, which determines whether or not the model is significant (i.e., whether any of the independent variables has an effect on the dependent variable). The test statistic's *F*-value is the result of a one-way ANOVA (see ▶ Chap. 6) that tests the null hypothesis that all the regression coefficients equal zero. Thus, the following null hypothesis is tested:[8]

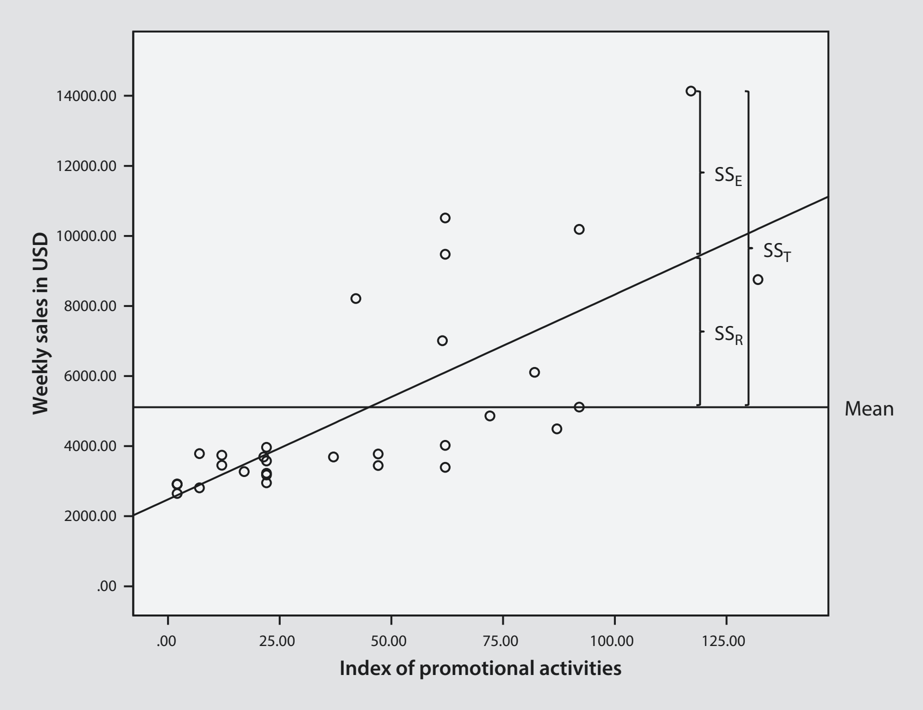
$$\mathrm{H}_0: \ \beta_1 = \beta_2 = \beta_3 = \ldots = 0$$

If the regression coefficients are all equal to zero, then all the independent variables' effect 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 non-significant. If the *p*-value of the *F*-test is below 0.05 (i.e., the model is significant), does not, however, imply that all the regression coefficients are significant, or even that one of them is significant when considered in isolation. However, if the *F*-value is significant, it is highly likely that at least one or more regression coefficients are significant. Note that model significance is not an indicator of (close) fit, only of significance.

If we find that the *F*-test is significant, we can interpret the model fit by using the $\boldsymbol{R^2}$. The R^2 (also called the **coefficient of determination**) indicates the degree to which the model, relative to the mean, explains the observed variation in the dependent variable. In ◘ Fig. 7.6, we illustrate this graphically by means of a scatter plot. The *y*-axis relates to the dependent variable *sales* (weekly sales in dollars) and the *x*-axis to the independent variable promotion (index of promotional activities). 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 $5000 sales per week) refers to the average sales across all 30 observations. This is also our benchmark. After all, if we were to have no regression line, our best estimate of the weekly sales would also be the average. The sum of all the squared differences between each observation and the average is the total variation or the **total sum of the squares** (**SS**$_\mathbf{T}$). We indicate the total variation in only one observation on the right of the scatter plot.

The straight upward sloping line (starting at the *y*-axis at about $2500 sales per week when there are no promotional activities) is the regression line that OLS estimates. If we want to understand what the regression model adds beyond the average (which is the benchmark for calculating the R^2), we can calculate the difference between the regression line and the line indicating the average. We call this the **regression sum of squares** (**SS**$_\mathbf{R}$), as it is the variation in the data that the regression analysis explains. The final point we need to understand regarding how well a regression line fits the available data, is the

[8] This hypothesis can also be read as that a model with only an intercept is sufficient.



■ **Fig. 7.6** Explanation of the R^2

unexplained sum of the squares. This is the difference between the observations (indicated by the dots) and the regression line. The squared sum of these differences refers to the regression error that we discussed previously and which is therefore denoted as the **error sum of squares** ($\mathrm{SS_E}$). 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 may see, this description is like the one-way ANOVA we discussed in ► Chap. 6. A useful regression line should explain a substantial amount of variation (have a high $\mathrm{SS_R}$) relative to the total variation ($\mathrm{SS_T}$). This degree of fit is indicated by the R^2 as follows:

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

The R^2 always lies between 0 and 1, with a higher R^2 indicating a better model fit. When interpreting the R^2, higher values indicate that the variation in x explains more of the variation in y. Therefore, relative to the $SS_{R,}$ the SS_E is low.

Tip

It is difficult to provide rules of thumb regarding what R^2 is appropriate, as this varies from research area to research area and on the model complexity. For example, in longitudinal studies, R^2s of 0.90 and higher are common. In cross-sectional designs, values of around 0.30 are common, while values of 0.10 are normal in cross-sectional data in exploratory research. In scholarly research focusing on marketing, R^2 values of 0.50, 0.30, and 0.10 can, as a rough rule of thumb, be respectively described as substantial, moderate, and weak. Similarly, the regression model's complexity influences the R^2 as the statistic increases with a greater number of independent variables.

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 only use the R^2 as the basis for understanding regression model fit, we are driven towards selecting regression models with many independent variables. Selecting a model only based on the R^2 is generally not a good strategy, unless we are interested in making predictions. If we are interested in determining whether independent variables have a significant relationship with a dependent variable, or when we wish to estimate the relative strength of different independent variables' effects, we need regression models that do a good job of explaining the data (which have a low SS_E), but which also have a few independent variables. It is easier to recommend that a management should change a few key variables to improve an outcome than to recommend a long list of somewhat related variables. We also do not want too many independent variables, because they are likely to complicate the insights. Consequently, it is best to rely on simple models when possible. Relevant variables should, of course, always be included. 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 the variance. We calculate the adjusted R^2 as follows:

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

Here, n describes the number of observations and k the number of independent variables (not counting the constant α). This adjusted R^2 is a relative measure and should be used to compare different but **nested models** with the same dependent variable. Nested means that all of a simpler model's terms are included in a more complex model, as well as additional variables. 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.

We cannot interpret the adjusted *R*² as the percentage of explained variance as we can with the regular *R*². The adjusted *R*² is only a measure of how much the model explains while controlling for model complexity.

Because the adjusted R^2 can only compare nested models, there are additional fit indices that can be used to compare models with the same dependent variable, but different independent variables (Treiman 2014). The **Akaike information criterion** (**AIC**) and the **Bayes information criterion** (**BIC**) are such measures of model fit. More precisely, AIC and BIC are a relative measure indicating the difference in information when a set of candidate models with different independent variables is estimated. For example, we can use these criteria to compare two models where the first regression model explains the *sales* by using two independent variables (e.g., *price* and *promotions*) and the second model adds one more independent variable (e.g., *price, promotions*, and *service quality*). We can also use the AIC and BIC when we explain sales by using two different sets of independent variables.

Both the AIC and BIC apply a penalty (the BIC a slightly larger one) as the number of independent variables increases with the sample size (Treiman 2014).[9] Smaller values are better and, when comparing models and a rough guide is that when the more complex model's AIC (or BIC) is 10 lower than that of another model, the former model should be given strong preference (Fabozzi et al. 2014). When the difference is less than 2, the simpler model is preferred. For values between 2 and 10, the evidence shifts towards the more complex model, although a specific cut-off point is hard to recommend. When interpreting these statistics, note that the AIC tends to point towards a more complex model than the BIC.

7.3.4.2 Effects of Individual Variables

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. If a regression coefficient's p-value is below 0.05, we generally say that the specific independent variable relates significantly to the dependent variable. To be precise, the null and alternative hypotheses tested for an individual parameter (e.g., β_1) are:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

If a coefficient is significant (i.e., the p-value is below 0.05), we reject the null hypothesis and support the alternative hypothesis, concluding that the parameter differs significantly from zero. For example, if we estimate a regression model on the data shown in ◘ Fig. 7.1, the (unstandardized) β_1 coefficient of promotional activities' effect on sales is 54.591, with a t-value of 5.354. This t-value results in a p-value less than 0.05, indicating that the effect is significantly different from zero. If we hypothesize a direction (i.e., smaller or larger than zero) instead of significantly different from zero, we should divide

[9] The AIC is specifically calculated as $AIC = \text{n}\left[\log\left(\frac{\text{SS}_\text{E}}{\text{n}}\right) + \frac{2\text{k}}{\text{n}}\right]$, where SS_E is the error sum of squares, *n* is the number of observations and *k* the number of independent variables, while the BIC is calculated as $BIC = \text{n}\left[\log\left(\frac{\text{SS}_\text{E}}{\text{n}}\right) + \frac{\text{k} \cdot \log(\text{n})}{\text{n}}\right]$. Note that these formulations only hold in case of normally distributed residuals with constant variance (Burnham and Anderson 2013).

the corresponding p-value by two. This is the same as applying the t-test for a directional effect, which is explained in ► Chap. 6.

The next step is to interpret the actual size of the β coefficients, which we can interpret in terms of **unstandardized effects** and **standardized effects**. The unstandardized β coefficient indicates the effect that a one-unit increase in the independent variable (on the scale used to measure the original independent variable) has on the dependent variable. This effect is therefore the partial relationship between a change in a single independent variable and the dependent variable. For example, the unstandardized β_1 coefficient of promotional activities' effect on sales (54.591) indicates that a one-unit change in (the index of) promotional activities increases sales by 54.591 units. Importantly, if we have multiple independent variables, a variable's unstandardized coefficient is the effect of that independent variable's increase by one unit, but 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 dollars, units sold, or on Likert scales. Consequently, the independent variables' effects cannot be directly compared with one another, because their influence also depends on the type of scale used. Comparing the unstandardized β coefficients would, in any case, amount to comparing apples with oranges!

Fortunately, the standardized βs allow us to compare the relative effect of differently measured independent variables by expressing the effect in terms of standard deviation changes from the mean. More precisely, the standardized β coefficient expresses the effect that a single standard deviation change in the independent variable has on the dependent variable. The standardized βs are used to compare different independent variables' effects. All we need to do is to find the highest absolute value, which indicates the variable that has the strongest effect on the dependent variable. The second highest absolute value indicates the second strongest effect, etc. Only consider significant βs in this respect, as insignificant βs do not (statistically) differ from zero! In practice, the standardized β is important, because it allows us to assess the effect of one variable (e.g., promotional activities), relative to other variables (e.g., prices). While the standardized βs are helpful from a practical point of view, they only allow for comparing the coefficients within and not between models! Even if you just add a single variable to your regression model, the standardized βs may change substantially.

When interpreting (standardized) β coefficients, you should always keep the effect size in mind. If a β coefficient is significant, it merely indicates an effect that differs from zero. This does not necessarily mean that the effect is managerially relevant. For example, we may find a \$0.01 sales effect of spending \$1 more on promotional activities that is statistically significant. Statistically, we could conclude that the effect of a \$1 increase in promotional activities increases sales by an average of \$0.01 (just one dollar cent). While this effect differs significantly from zero, we would probably not recommend increasing promotional activities in practice (we would lose money on the margin) as the effect size is just too small.[10]

[10] Cohen's (1994) classical article "The Earth is Round (p < .05)" offers an interesting discussion on significance and effect sizes.

Box 7.2 Moderation analysis

The discussion of individual variables' effects assumes that there is only one effect. That is, that only one β parameter represents all observations well. This is often not true. For example, the link between sales and price has been shown to be stronger when promotional activities are higher. In other words, the effect of price (β_1) is not constant, but varies with the level of promotional activities.

Moderation analysis is one way of testing if such heterogeneity is present. A moderator variable, usually denoted by *m*, is a variable that changes the strength (or even direction) of the relationship between the independent variable (x_1) and the dependent variable (y). You only need to create a new variable that is the multiplication of x_1 and *m* (i.e., $x_1 \cdot m$). The regression model then takes the following form:

$$y = \alpha + \beta_1 x_1 + \beta_2 m + \beta_3 x_1 \cdot m + e$$

In words, a moderator analysis requires entering the independent variable $x_{1,}$ the moderator variable *m*, and the product $x_1 \cdot m$ (commonly referred to as *interaction term*), which represents the interaction between the independent variable and the moderator. Moderation analysis is therefore also commonly referred to as an analysis of *interaction effects*. After estimating this regression model, you can interpret the significance and sign of the β_3 parameter. A significant effect suggests that

- when the sign of β_3 is positive, the effect β_1 increases as *m* increases,
- when the sign of β_3 is negative, the effect β_1 decreases as *m* increases.

For further details on moderation analysis, please see David Kenny's discussion on moderation (http://www.davidakenny.net/cm/moderation.htm), or the advanced discussion by Aiken and West (1991). Jeremy Dawson's website (http://www.jeremydawson.co.uk/slopes.htm) offers a tool for visualizing moderation effects. An example of a moderation analysis is found in Mooi and Frambach (2009).

There are also situations in which an effect is not constant for all observations, but depends on another variable's values. Researchers can run a **moderation analysis**, which we discuss in Box 7.2, to estimate such effects.

7.3.5 Validate the Regression Results

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

1. We can randomly split the dataset into two parts (called **split-sample validation**) and run the regression model again on each data subset. 70 % of the randomly chosen data is often used to estimate the regression model (called **estimation sample**) and the remaining 30 % is used for comparison purposes (called **validation sample**). 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. 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 $p = 0.051$)..[11] Finally, note that it is mere convention to use 70 and 30 % and there is no specific reason for using these percentages.

2. We can also cross-validate our findings on a new dataset and examine whether these findings are similar to the original findings. Again, similarity in the findings indicates stability and that our regression model is properly specified. **Cross-validation** does, of course, assume that we have a second dataset.
3. We can add several alternative variables to the model and examine whether the original effects change. For example, if we try to explain weekly supermarket sales, we could use several additional variables, such as the breadth of the assortment or the downtown/non-downtown location in our regression model. If the basic findings we obtained earlier continue to hold even when adding these two new variables, we conclude that the effects are stable. This analysis does, of course, require us to have more variables available than those included in the original regression model. If we add variables not because we are interested in them but because we want to rule these out as alternative explanations, we call these **control variables**.

7.3.6 Use the Regression Model

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

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

If we estimate this model on the previously used dataset, the estimated unstandardized coefficients using regression analysis are 30,304.054 for the intercept, –25,209.858 for price, and 42.266 for promotions. We can use these coefficients to predict sales in different situations. Imagine, for example, that we set the price at \$1.10 and the promotional activities at 50. Our expectation of the weekly sales would then be:

$$\hat{y} = 30{,}304.054 - 25{,}209.858 \cdot \$1.10 + 42.266 \cdot 50 \text{ promotional activities} = \$4686.5102 \text{ sales per week.}$$

[11] It is possible to compare regression coefficients statistically, avoiding the need to the subjectivity of "similar." Strictly speaking, the test for comparing coefficients is z-distributed with $z = \frac{b_1 - b_2}{\sqrt{SE_1^2 + SE_2^2}}$ (see Paternoster et al. 1998).

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

Regression can also help by providing insight into variables' specific effects. For example, if the effect of promotions is not significant, it may tell managers that the supermarket's sales are insensitive to promotions. Alternatively, if there is some effect, the strength and direction of promotional activities' effect may help managers understand whether they are useful.

◘ 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.

◘ **Table 7.2** Steps involved in carrying out a regression analysis

Theory	Action
Check the regression analysis data requirements	
Sufficient sample size	Run a power analysis using G*power. Check if sample size is 104 + k, where k indicates the number of independent variables. If the expected effects are weak (the R^2 is 0.10 or lower), use at least 30 · k observations per independent variable. To check the sample size, go to ► Analyze ► Regression ► Linear ► Statistics and check **Descriptives**. In the output check **N** under **Descriptive Statistics**. Alternatively, after running a regression analysis, look at the intersection of **df** and **Total** in the **ANOVA** table. You need to add 1 to this number to get the total sample size.
Do the dependent and independent variables show variation?	Calculate the standard deviation of the variables by going to ► Analyze ► Regression ► Linear ► Statistics and check **Descriptive**. In the output check **Std. Deviation** under **Descriptive Statistics**. At the very least, the standard deviation should be greater than 0.
Is the dependent variable interval or ratio scaled?	See ► Chap. 3 determine the measurement level.
Is (multi)collinearity present?	Check the VIF. Go to ► Analyze ► Regression ► Linear ► Statistics and check **Collinearity diagnostics**. In the output, under **Collinearity Statistics**, assess if the VIFs are all below 10 (although the threshold can be higher, or lower, in some cases; see ► Sect. 7.3.1.4 for details). Note that SPSS also produces an additional table labelled **Collinearity Diagnostics** that is not needed.

Table 7.2 (Continued)

Theory	**Action**
Specify and estimate the regression model	
Model specification	1. Pick distinct variables 2. Try to build a robust model 3. Consider the variables that are needed to give advice 4. Consider whether the number of independent variables is in relation to the sample size When using ordinal variables, create dummies first. Do this by going to ► Transform ► Create Dummy Variables. Move the ordinal variable to the **Create Dummy Variables for** box and also enter the first part of the name of the dummy under **Root Names (One per Selected Variable)** of the **Main Effect Dummy Variables box**. By default, SPSS chooses to give value labels to each category. We can change this to **Use values** if we have not defined value labels. Note that SPSS will automatically create dummies for all categories and we have to specify the base category.
Estimate the regression model	► Analyze ► Regression ► Linear. Under **Dependent**, enter the dependent variable and add all the independent variables under **Independent(s)**. Click on **OK**.
Test the regression analysis assumptions	
Can the regression model be specified linearly?	Consider whether you can write the regression model as: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + e$
Is the relationship between the independent and dependent variables linear?	Plot the dependent variable against the independent variable using a scatterplot matrix to see if the relation (if any) appears to be linear. ► Graphs ► Chart Builder. Then click on **Scatter/Dot** and drag **Simple Scatter** to the **Chart preview** window. Drag the dependent variable onto the **Y-axis** and the first independent variable to the **X-axis**. Repeat for the other independent variables. To conduct Ramsey's RESET test first run the regression model by going to ► Analyze ► Regression ► Linear. Under **Dependent**, enter the dependent variable and add all the independent variables under **Block 1 of 1** and click on **Save**. Then click on **Unstandardized** under **Predicted Values**, followed by **Continue** and **OK**. After this go to ► Transform ► Compute Variable. Under **Target Variable**, type *PRED_2* and under **numeric expression** enter the unstandardized prediction variable *PRE_1* and type ****2** after this to compute the squared predicted values. Repeat this for predicted values to the power of 3 by entering *PRED_3* under **Target Variable**. Under **numeric expression** enter the unstandardized prediction variable *PRE_1* and type ****3**. Then re-run the regression model by going to ► Analyze ► Regression ► Linear. Under **Dependent**, enter the dependent variable and add all the independent variables under **Block 1 of 1.** Then click on **Next** and enter *PRED_2* and *PRED_3* to the box. Under **Statistics** tick **R squared change** and click on **Continue** and then **OK**. Check under **Model Summary** if **Sig. F Change** is significant for model **2**.

■ **Table 7.2** (Continued)

Theory	**Action**
Is the expected mean error of the regression model zero?	Choice made on theoretical grounds.
Is the error variance constant (homoscedasticity)?	Re-run the regression model by going to ► Analyze ► Regression ► Linear. Under **Dependent**, enter the dependent variable and add all the independent variables under **Block 1 of 1** and click on **Save**. Then go to **Save** and click on **Unstandardized** under **Residuals**. Then click on **Continue** and **OK**. Plot the dependent variable against the residual by going to ► Graphs ► Chart Builder. Then click on **Scatter/Dot** and drag **Simple Scatter** to the **Chart preview** window. Drag the dependent variable onto the **Y-axis** and the saved residual onto the **X-axis.** Assess if there is a clear in- or decrease in the error variance (i.e. a funnel shape). If there is evidence of heteroscedasticity we can use bootstrapping. Re-run the regression model by going to ► Analyze ► Regression ► Linear and click on **Bootstrap**. Then click on **Perform bootstrapping** and make sure that **Number of samples** is set to 1,000. Click on **Continue** and **OK**. Note that this requires the SPSS bootstrapping add-on. If this is not installed, you will not see this option. Check if the interpretation of the regression results is the same for the standard as for the bootstrapped significance (under **Sig. (2-tailed)**). Under the column **Bias**, you can see the difference with the normal *ß*s. The bias should be small. That is, bootstrapped parameter estimates should fall in the region of 2 ± standard errors from the coefficients of the model estimated via OLS. Compare the normal and bootstrapped model results to ensure they are similar.
Are the errors independent (no autocorrelation)?	First assess if there is a time component to the data (i.e., multiple observations, across time, from one respondent/object). If there is, sort the data according to the time variable and conduct the Durbin–Watson test. Compare the calculated Durbin–Watson test statistic with the critical lower and upper values. If positive or negative autocorrelation is present, panel or time-series models need to be used. Re-run the regression model by going to ► Analyze ► Regression ► Linear. Under **Dependent**, enter the dependent variable and add all the independent variables under **Block 1 of 1** and click on **Statistics**. Then click on **Durbin-Watson** under **Residuals**, followed by **Continue** and **OK**. The Durbin-Watson test for first-order serial correlation should not be significant. The critical values can be found on the website accompanying this book (↓ Web Appendix → Downloads).

Table 7.2 (Continued)

Theory	Action
Are the errors approximately normally distributed?	Visual inspection: Re-run the regression model by going to ► Analyze ► Regression ► Linear. Under **Dependent**, enter the dependent variable and add all the independent variables under **Block 1 of 1** and click on **Plots**. Then click on **Histogram** under **Standardized Residual Plots**. Then click on **Continue** and **OK**. Check whether the length of the bars of the histogram follows to the normal curve. Statistical test: Re-run the regression model by going to ► Analyze ► Regression ► Linear. Under **Dependent**, enter the dependent variable and add all the independent variables under **Block 1 of 1** and click on **Save**. Then click on **Unstandardized** under **Residuals**. Then click on **Continue** and **OK**. Calculate the Shapiro–Wilk test. ► Analyze ► Descriptive Statistics ► Explore. Add the unstandardized residual to the **Dependent list** and click on **Plots**. Tick the **Normality plots with tests** box and click on **Continue** and **OK**. Check if the Shapiro-Wilk test under **Sig.** reports a *p*-value greater than 0.05.
Interpret the regression model	
Consider the overall model fit	Check the significance of the *F*-test and the R^2 value.
Consider the effects of the independent variables separately	Check the (standardized) β. Also check the sign of the β. Consider the significance of each coefficient by checking the *p*-value und **Sig.**.
To compare models	Calculate the AIC and BIC. This can only be done by setting up the regression model and instead of clicking on **OK**, click on **Paste.** Add **SELECTION** to the/**STATISTICS** subcommand (e.g./ **STATISTICS COEFF OUTS R ANOVA COLLIN TOL SELECTION**). Select the entire code and go to ► Run ► All. Check the AIC and BIC and ascertain if the simpler model has AIC and BIC values that are at least 2, but preferably 10, lower than that of the more complex model.
Validate the model	
Are the results robust?	Randomly select 30 % of cases (assuring minimum sample size requirements apply). To do this, go to ► Data ► Select Cases. Then click on **Random sample of cases**, followed by **Sample**. Type **30** in the box **Approximately … % of all cases** and click on **Continue**. Before clicking on **OK**, make sure to select **Filter out unselected cases** under **Output**. Now click on **OK**. Then go to ► Data ► Select Cases and select **All cases**. Next, go to ► Data ► Split File. In the dialog box that opens, select the option **Organize output by groups**, move the *filter_$* variable into the **Groups Based on** box, and click on **OK**. Compare the model results to ensure they are similar (i.e., the coefficients of the regression run on the smaller sample are similar to the coefficients of the regression model run on the larger sample in terms of direction and significance.

7.4 Example

Let's go back to the Oddjob Airways case study (as introduced in ► Chap. 5) and run a regression analysis on the data. Our aim is to explain commitment, which is the customer's intention to continue the relationship. This variable is the mean of the three items *com1* ("I am very committed to Oddjob Airways"), *com2* ("My relationship with Oddjob Airways means a lot to me"), and *com3* ("If Oddjob Airways would not exist any longer, it would be a hard loss for me") and has already been included in the dataset (↓ Web Appendix → Downloads).

Our task is to identify which variables relate to commitment to Oddjob Airways. Regression analysis can help us determine which variables relate significantly to commitment, while also identifying the relative strength of the different independent variables. The Oddjob Airways dataset offers several variables that may explain the *commitment* variable. Based on prior research and discussions with Oddjob Airway's management, the following variables have been identified as promising candidates:

- Oddjob Airways gives you a sense of safety (*s9*),
- The condition of Oddjob Airways' aircraft is immaculate (*s10*),
- Oddjob Airways also pays attention to its service delivery's details (*s19*),
- Oddjob Airways makes traveling uncomplicated (*s21*), and
- Oddjob Airways offers great value for money (*s23*).

As additional variables, we add the following three categories to the model: the respondent's status (*status*), age (*age*), and gender (*gender*).

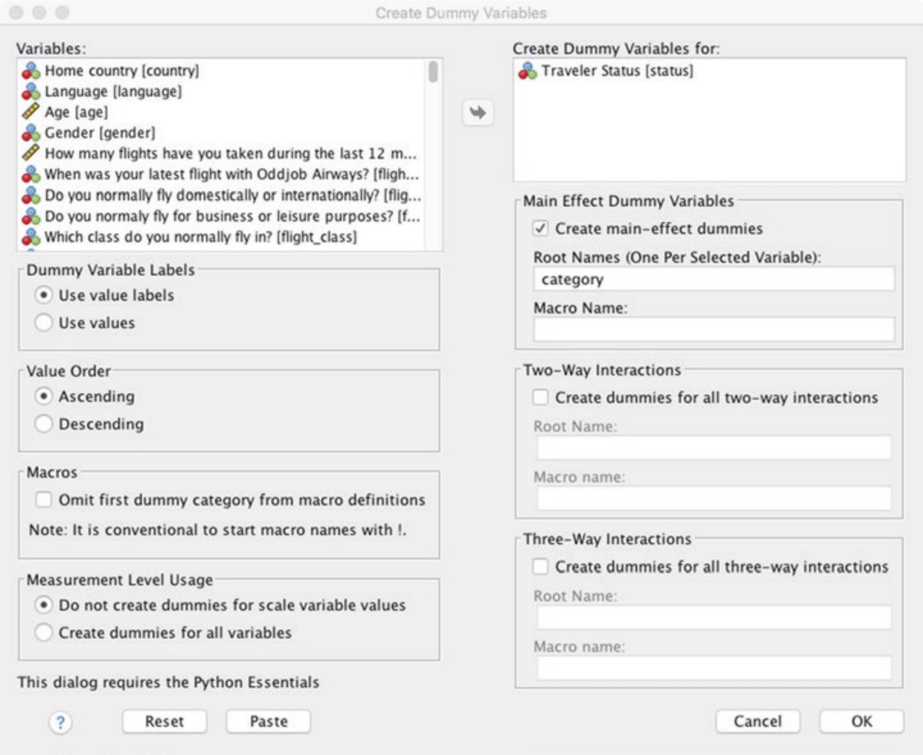
■ Check the Regression Analysis Data Requirements

To check whether our data meet all requirements, we need to run the regression analysis first. Before we run a regression analysis, however, we need to create dummy variables from the *status* variable. To do this, go to ► Transform ► Create Dummy Variables. Move the ordinal variable *status* into the **Create Dummy Variables for** box and also enter the first part of the name of the dummy as *category* under **Root Names (One per Selected Variable)** of the **Main Effect Dummy Variables** box (◘ Fig. 7.7). By default, SPSS chooses to give value labels to label each category which is good since value labels have been defined as *Blue, Silver*, and *Gold*. Then click on **OK.**[12]

Next, we should run the regression analysis. Go to ► Analyze ► Regression ► Linear and enter the dependent variable *commitment* in the **Dependent** box and *s9, s10, s19, s21, s23, category_2, category_3, age*, and *gender* in the **Independent(s)** box as shown in ◘ Fig. 7.8. This will require a bit of scrolling. Note that if we were to add *category_1* as well, this variable would be perfectly collinear with *category_2* and *category_3* since if we know an observation is not in the latter two, it must be in the former. SPSS issues no warning when it drops a variable.

Next, we need to tell SPSS that we require several options to conduct the steps as discussed earlier. First, click on **Statistics** and check (if it isn't already checked) **Estimates,**

[12] Note that this only works, as shown in the lower left of ◘ Fig. 7.7, if the "Python essentials" are installed.



Create Dummy Variables

Variables:
- Home country [country]
- Language [language]
- Age [age]
- Gender [gender]
- How many flights have you taken during the last 12 m...
- When was your latest flight with Oddjob Airways? [fligh...
- Do you normally fly domestically or internationally? [flig...
- Do you normaly fly for business or leisure purposes? [f...
- Which class do you normally fly in? [flight_class]

Dummy Variable Labels
- ● Use value labels
- ○ Use values

Value Order
- ● Ascending
- ○ Descending

Macros
- □ Omit first dummy category from macro definitions

Note: It is conventional to start macro names with !.

Measurement Level Usage
- ● Do not create dummies for scale variable values
- ○ Create dummies for all variables

This dialog requires the Python Essentials

Create Dummy Variables for:
- Traveler Status [status]

Main Effect Dummy Variables
- ☑ Create main-effect dummies

Root Names (One Per Selected Variable): category

Macro Name:

Two-Way Interactions
- □ Create dummies for all two-way interactions

Root Name:

Macro name:

Three-Way Interactions
- □ Create dummies for all three-way interactions

Root Name:

Macro name:

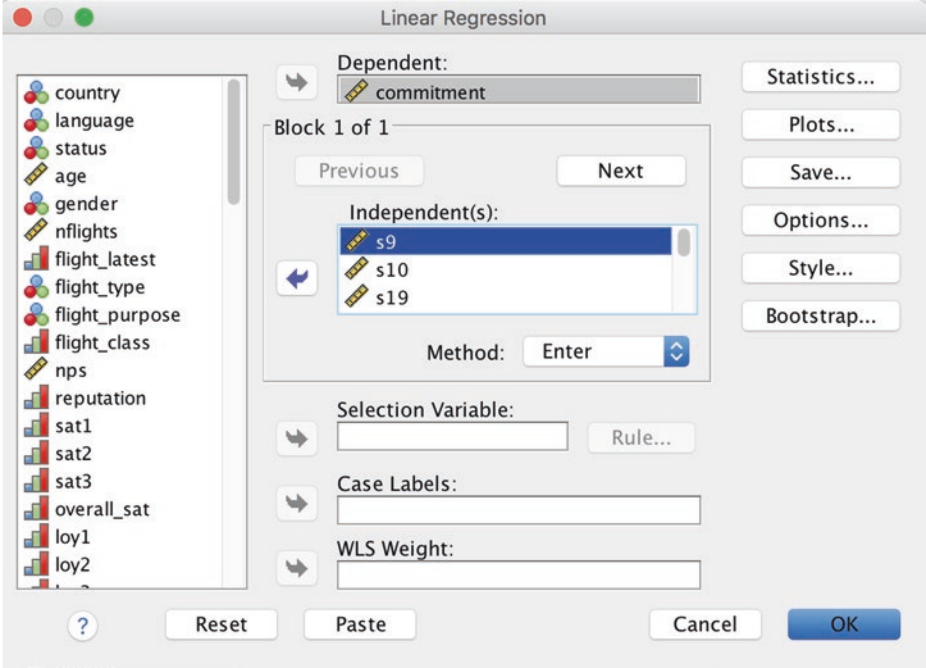
■ **Fig. 7.7** The create dummy variables dialog box

Model fit, Descriptives, and **Collinearity diagnostics** (■ Fig. 7.9). You should also check Durbin-Watson if the data have a time component and are sorted based on this component. In these data, there is no time component so this should not be checked. Then click on **Continue**.

To save the predicted values and the errors we should click on **Save**. Select the option **Unstandardized** right under **Predicted Values** and under **Residuals** (■ Fig. 7.10). Then click on **Continue**, followed by **OK**, which will initiate the regression analysis.

Sample Size

After having run the analysis, we should first check if the sample size is sufficient for a regression analysis. To do so, we examine the **Descriptive Statistics** table (■ Table 7.3), which shows the mean, the standard deviation, and the number of observations (indicated by **N**). Here, we find the value **973**. Green's (1991) rule of thumb suggests that we need at least 104 + *k* observations, where *k* is the number of independent variables. Since we have nine independent variables, we satisfy this criterion. In fact, even if we apply VanVoorhis and Morgan's (2007) more stringent criteria of 30 observations per variable, we still have a sufficient sample size. SPSS also displays a correlation matrix. However, as this matrix takes up a lot of space, we don't display it here.

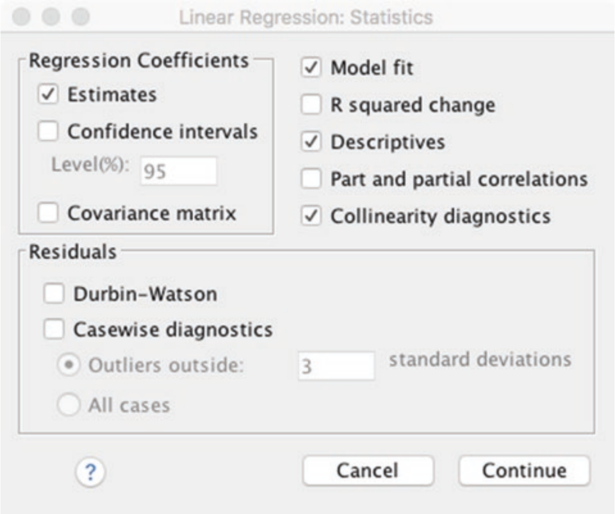


Dependent: commitment

Independent(s): s9, s10, s19

Method: Enter

Fig. 7.8 The regression dialog box

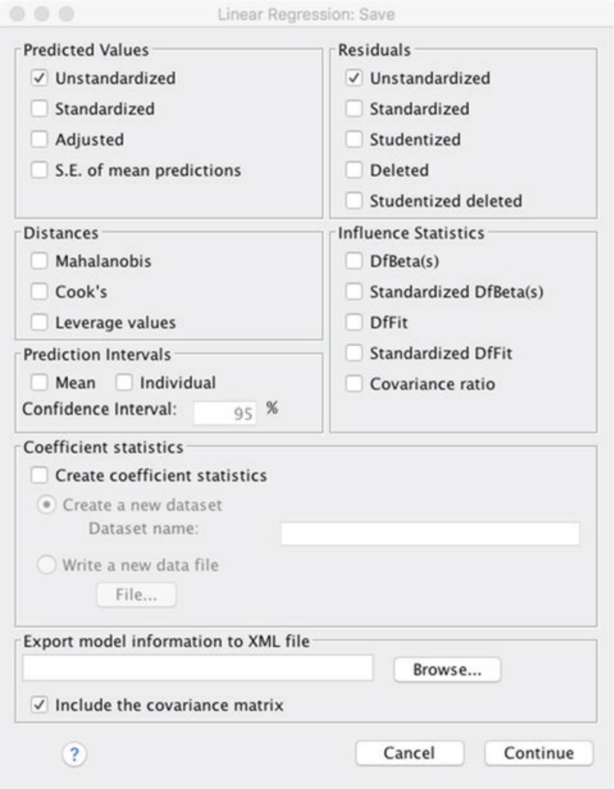


☑ Estimates
□ Confidence intervals
Level(%): 95
□ Covariance matrix
☑ Model fit
□ R squared change
☑ Descriptives
□ Part and partial correlations
☑ Collinearity diagnostics
□ Durbin-Watson
□ Casewise diagnostics
● Outliers outside: 3 standard deviations
○ All cases

Fig. 7.9 The statistics option (regression analysis)

Variables Need to Vary

To ascertain if our variables display some variation, we examine the **Descriptive Statistics** output from Table 7.3 again. At the very least, the standard deviation, indicated under



Linear Regression: Save

Predicted Values
☑ Unstandardized
□ Standardized
□ Adjusted
□ S.E. of mean predictions

Residuals
☑ Unstandardized
□ Standardized
□ Studentized
□ Deleted
□ Studentized deleted

Distances
□ Mahalanobis
□ Cook's
□ Leverage values

Influence Statistics
□ DfBeta(s)
□ Standardized DfBeta(s)
□ DfFit
□ Standardized DfFit
□ Covariance ratio

Prediction Intervals
□ Mean □ Individual
Confidence Interval: 95 %

Coefficient statistics
□ Create coefficient statistics
● Create a new dataset
Dataset name:
○ Write a new data file

Export model information to XML file
☑ Include the covariance matrix

Fig. 7.10 The save option (regression analysis)

Table 7.3 Descriptive statistics

Descriptive Statistics

	Mean	Std. Deviation	N
commitment	4.2254	1.71537	973
s9	72.63	20.795	973
s10	64.65	21.370	973
s19	57.41	21.489	973
s21	59.18	22.535	973
s23	49.37	22.675	973
category_2	.2343	.42380	973
category_3	.1377	.34478	973
age	50.58	11.959	973
gender	1.74	.440	973

Std. Deviation should be greater than 0. The output also shows each variable's mean. Two variables require a bit more to interpret these. The variable *status* is coded using thee labels, *Blue, Silver*, and *Gold*, and from the output we can see that **.2343** or 23 % of all respondents fall into *category_2* (*Silver*) category, while **.1377** or 14 % fall into *category_3* (*Gold*). Hence, we can conclude that the reaming 63 % fall into the *Blue* category. Turning to the variable *gender*, which is coded 1 for females and 2 for males, we can conclude that as the mean is 1.74, 74 % of our observations are male.

Scale Type of the Dependent Variable

The scale of the dependent variable is interval or ratio scaled. Specifically, three 7-point Likert scales create the mean of three items that form the *commitment* variable. Most researchers would consider this to be interval or ratio scaled, which meets the OLS regression data assumptions.

Collinearity

To check for collinearity among the independent variables, we need to examine the regression output (Table 7.4), which shows the beta coefficients, their significances, and collinearity statistics. As we can see, the highest VIF value is **2.069**, which is below 10 and also below the conservative threshold of 5. Hence, we conclude that collinearity is not at a critical level. Note that SPSS will also produce a specific table with collinearity diagnostics but this table is not necessary for an essential diagnostic of collinearity.

Table 7.4 Regression output

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Toler-ance	VIF
1	(Constant)	1.199	.300		3.997	.000		
	s9	.005	.003	.063	1.722	.085	.521	1.919
	s10	.001	.003	.008	.224	.823	.498	2.009
	s19	.012	.003	.154	4.072	.000	.483	2.069
	s21	.019	.003	.245	6.821	.000	.532	1.880
	s23	.016	.003	.208	6.003	.000	.571	1.752
	category_2	.183	.112	.045	1.641	.101	.902	1.108
	category_3	.428	.138	.086	3.105	.002	.897	1.115
	age	.010	.004	.072	2.667	.008	.951	1.051
	gender	–.345	.105	–.089	–3.285	.001	.946	1.057

[a] Dependent Variable: commitment

Having now met all the described requirements for a regression analysis, our next task is to interpret the regression analysis results.

Specify and Estimate the Regression Model

We know exactly which variables to select for this model: *commitment*, as the dependent variable, and *s9, s10, s19, s21, s23, category_2, category_3, age*, and *gender* as the independent variables as we had specified earlier. We have already done this so there is no need to repeat this.

Test the Regression Analysis Assumptions

In the next step, we test the following assumptions relevant to regression analysis:

- Linearity,
- expected mean is zero,
- homoscedasticity,
- no autocorrelation, and
- normal error distribution.

First Assumption: Linearity

The first assumption is whether the regression model can be expressed linearly. Since no variable transformations occurred, we meet this assumption, because we can write the regression model linearly as:

$$\begin{aligned} commitment = {} & \alpha + \beta_1 s9 + \beta_2 s10 + \beta_3 s19 + \beta_4 s21 + \beta_5 s23 + \\ & \beta_6 category\_2 + \beta_7 category\_3 + \beta_8 age + \beta_9 gender + e \end{aligned}$$

Separately, we also check whether the relationships between the independent and dependent variables are linear. To do this, create scatterplots of the dependent variable against all the independent variables. To do this, go to ► Graphs ► Chart Builder. Then click on **Scatter/Dot** and drag **Simple Scatter** to the **Chart preview** window. Drag the dependent variable *commitment* onto the **Y-Axis** and the first independent variable *s9* ("Oddjob Airways gives you a sense of safety") to the **X-Axis** and click on **OK**.

▣ Figure 7.11 shows the resulting scatterplot. The large number of dots makes it difficult to see whether the relationship is linear and this is quite typical for large datasets. The scatterplot neither suggests nor rejects linearity clearly. Note that because *commitment* is computed as the mean of three variables measured on seven-point scales, we observe distinct bands. This is because the variable *commitment* can only take on 21 distinct values.

Another means to test for nonlinear relationships between the independent and dependent variable is to run Ramsey's RESET test. Unfortunately, the test is not readily available via SPSS's graphical user interface but we can manually run the test using a series of commands. To conduct Ramsey's RESET we need to square and take the third power of the predictions of the regression model we saved earlier. To do this, go to ► Transform ► Compute Variable and type *PRED_2* under **Target Variable**. Under **Numeric Expression** enter the unstandardized prediction variable *PRE_1* that we requested SPSS to save when running the regression analysis. Now type ****2** after *PRE_1* to compute the squared predicted values. After clicking on **OK**, SPSS will add a new variable *PRED_2* to the dataset. Repeat this series of commands but enter *PRED_3* under **Target Variable** and add ****3** after *PRE_1* to raise to the power of three.

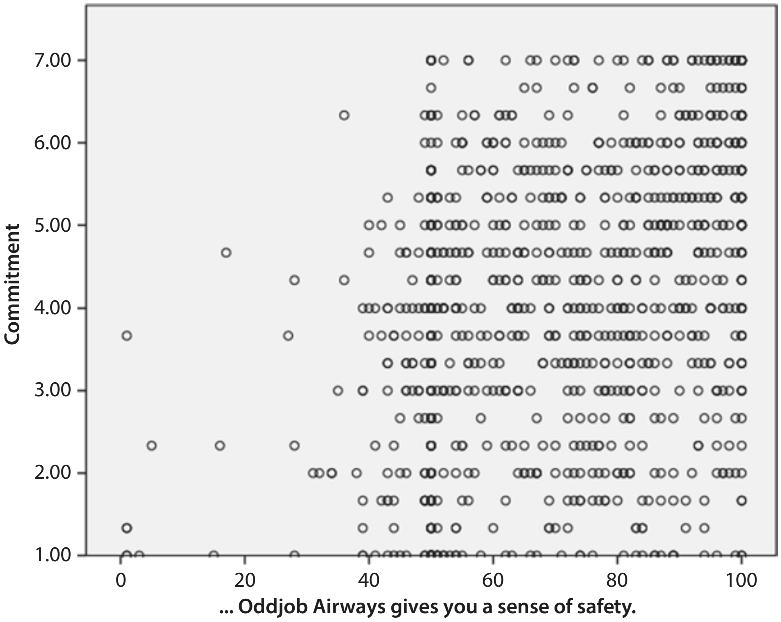


Fig. 7.11 Scatterplot to examine linearity

Then re-run the regression model by going to ► Analyze ► Regression ► Linear Regression. Under **Dependent** enter the dependent variable *commitment* and add all the independent variables (i.e., and *s9, s10, s19, s21, s23, category_2, category_3, age*, and *gender*) under **Block 1 of 1.** Then click on **Next** and enter *PRED_2* and *PRED_3* to the box. Note that this block will now read **Block 2 of 2**. Under **Statistics**, tick **R squared change** and click on **Continue** and then **OK**.

SPSS produces a series of outputs of which only the **Model Summary** is relevant to us. Specifically, we need to check whether the R^2 differs significantly between model 1 (the initial model) and model 2 (the model additionally containing *PRED_2* and *PRED_3*). If we read the output in Table 7.5, we see that the change in R^2 is not significant as indicated by the *p*-value of **.800** under **Sig. F Change**. The results thus suggest that the relationships are linear. Bear in mind, however, that this test does not consider all forms of non-linearities.

Table 7.5 Model summary used for Ramsey's RESET test

Model Summary[c]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.581[a]	.338	.332	1.40236	.338	54.593	9	963	.000
2	.582[b]	.338	.331	1.40349	.000	.223	2	961	.800

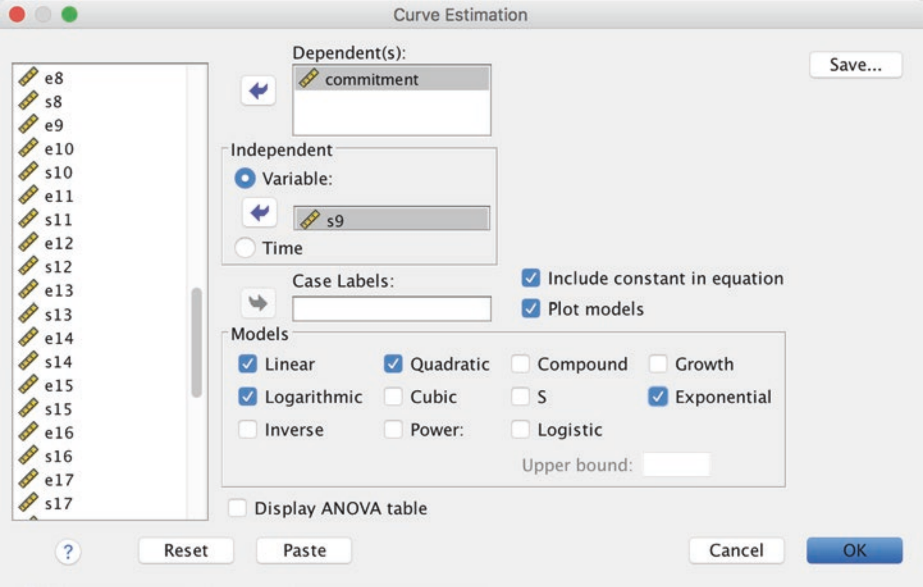
[a] Predictors: (Constant), gender, s21, age, category_3, category_2, s10, s23, s9, s19
[b] Predictors: (Constant), gender, s21, age, category_3, category_2, s10, s23, s9, s19, PRED_3, PRED_2
[c] Dependent Variable: commitment

In case the visual inspection or Ramsey's RESET test indicate a nonlinear relationship, we can run a curve estimation to explore the concrete form of the relationship. In Box 7.3, we offer a brief introduction into curve estimation using SPSS.

Box 7.3 Curve estimation

SPSS's curve estimation option allows running multiple regression models with a single independent variable, assuming different types of nonlinear relationships. To run this analysis, go to ► Analyze ► Regression ► Curve Estimation. In the dialog box that opens (◘ Fig. 7.12), enter *commitment* into the **Dependent(s)** box and, for example, *s9* under **Independent**. Under **Models** we can select different types of transformations to map a potential nonlinear relationship. Typical transformations include the **Logarithmic, Quadratic**, and **Exponential**. Make sure to also select **Linear**. Now click on **OK**.

◘ Figure 7.13 shows the resulting output with linear, logarithmic, quadratic, and exponential transformations. Considering the great number of data points, it is difficult to spot, which type of transformation would better represent the relationship between *commitment* and *s9*. In this case, we should compare the R^2 of all significant models as an indicator of what line fits best. To do so, select **Display ANOVA table** option in ◘ Fig. 7.12 and pick the model with a significant *F*-value, which has the highest R^2.[13]



Dependent(s): commitment

Independent
● Variable: s9
○ Time

Case Labels:

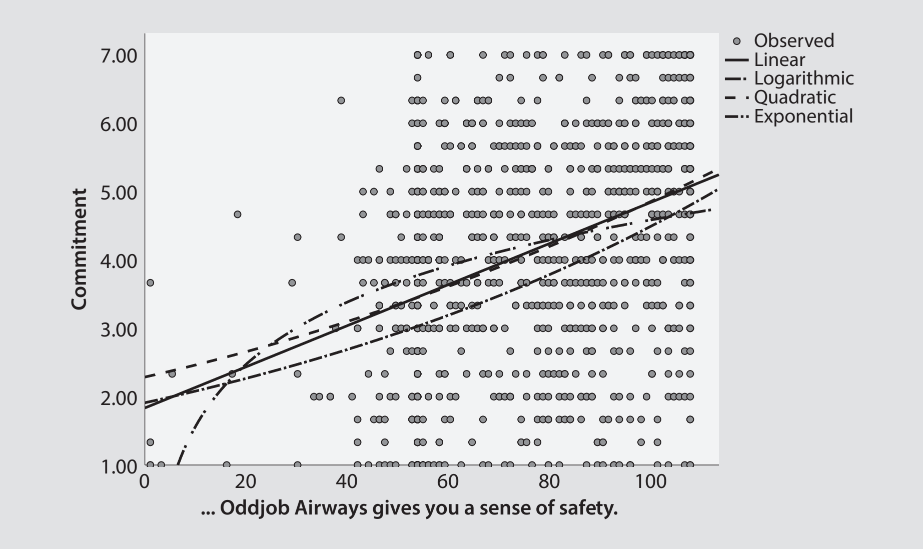
☑ Include constant in equation
☑ Plot models

Models
☑ Linear ☑ Quadratic □ Compound □ Growth
☑ Logarithmic □ Cubic □ S ☑ Exponential
□ Inverse □ Power: □ Logistic
Upper bound:

□ Display ANOVA table

◘ **Fig. 7.12** Curve estimation dialog box

[13] Note that it is better to calculate if the R^2 increase is significant (as for Ramsey's RESET test) but this needs to be done manually and falls outside of the scope of this book.



■ **Fig. 7.13** Examples of linear, logarithmic, quadratic, and exponential relationships

Second Assumption: Expected Mean Error Is Zero

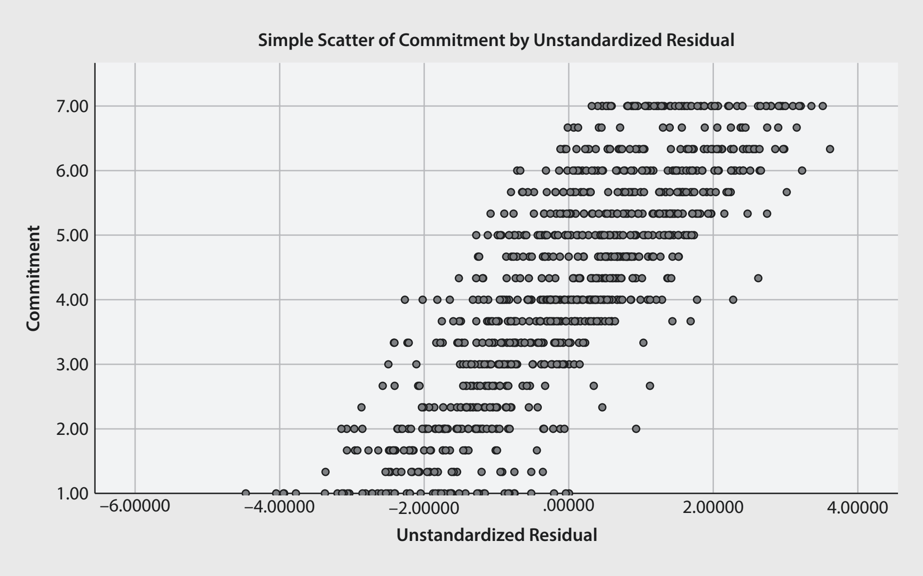
Assessing whether the regression model's expected mean error is zero is made on theoretical grounds—there is no empirical test for this. We have a randomly drawn sample from the population and the model is similar in specification to other models explaining commitment. This makes it reasonable to assume that the regression model's expected mean error is zero.

Third Assumption: Homoscedasticity

To consider if the errors are constant (homoscedastic) we need to re-use the errors of the regression model we saved earlier. We should then plot the dependent variable against the residual by going to ► Graphs ► Chart Builder. Then click on **Scatter/Dot** and drag **Simple Scatter** to the **Chart preview** window. Drag the dependent variable *commitment* onto the **Y-axis** and the unstandardized residuals variable *RES_1* onto the **X-axis**. Click on **OK**. The resulting scatterplot shown in ■ Fig. 7.14 shows no clear increase or decrease in error variance (i.e., a funnel shape). Since there is no clear evidence of heteroscedasticity, we do not need to use bootstrapping.

Fourth Assumption: No Autocorrelation

If we had data with a time component, we would also perform the Durbin-Watson test to check for potential autocorrelation. This requires us to first specify a time component, which is absent in the Oddjob Airways dataset.



■ **Fig. 7.14** Scatterplot to assess homoscedasticity

Fifth (Optional) Assumption: Normal Error Distribution

To check whether the errors are normally distributed, we need to re-run the regression model by going to ► Analyze ► Regression ► Linear Regression. Under **Dependent**, enter the dependent variable *commitment* and add all the independent variables (i.e., *s9, s10, s19, s21, s23, category_2, category_3, age*, and *gender*) under **Block 1 of 1** (in case you ran Ramsey's RESET test, you should press **RESET** first). Next, click on **Plots**. In the dialog box that opens, click on **Histogram** under **Standardized Residual Plots**, followed by **Continue** and **OK**. SPSS will then produce a plot as shown in ■ Fig. 7.15, which 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 Shapiro–Wilk test. Do this by going to ► Analyze ► Descriptive Statistics ► Explore. Add the unstandardized residual variable *RES_1* to the **Dependent list** box and click on **Plots**. Tick the **Normality plots with tests** box and click on **Continue** and **OK**. Then check if the Shapiro-Wilk test under **Sig.** reports a greater value than 0.05. We see this is the case (**.083**) and therefore conclude that the residual is normally distributed (■ Table 7.6).

Interpret the Regression Model

We have already conducted a regression analysis previously to test the assumptions, so there's no need to run the analysis again. We only need to interpret ■ Table 7.4 in detail and the two tables that precede this regression table and which we haven't interpreted previously. These are ■ Tables 7.7 and 7.8.

We start our interpretation of the regression model by examining the model fit. To do so, we first check if the model is significant by interpreting the *F*-test. The *p*-value under **Sig.**

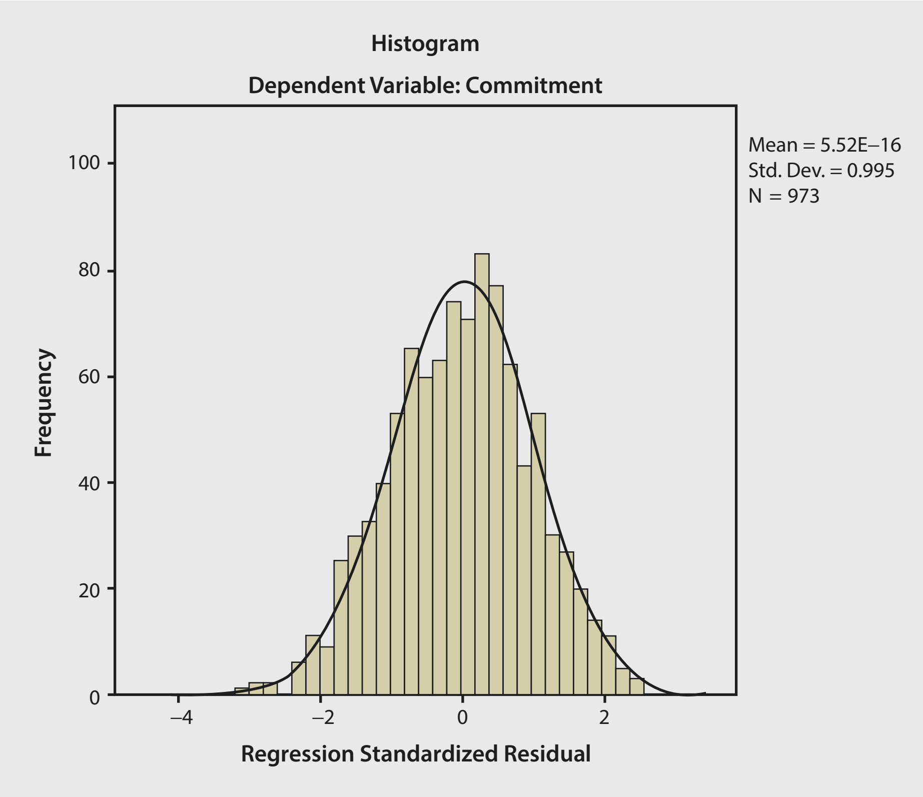


Fig. 7.15 Histogram of the errors with a standard normal curve

Table 7.6 Tests of normality

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RES_1	.026	973	.113	.997	973	.083

[a] Lilliefors Significance Correction

in Table 7.8 (**.000**) is clearly lower than 0.05, indicating that the model is significant.[14] In the next step, we interpret the model's R^2, which is given in Table 7.7. The value of **.338** is moderate but still satisfactory, considering that customer commitment is a rather abstract concept, which is therefore difficult to predict. The adjusted R^2 is **.332** but this value is only useful when comparing models with different predictors.

[14] Note that a *p*-value is never exactly zero, but has values different from zero in later decimal places.

Table 7.7 Model summary

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.581[a]	.338	.332	1.40236

[a] Predictors: (Constant), gender, s21, age, category_3, category_2, s10, s23, s9, s19
[b] Dependent Variable: commitment

Table 7.8 ANOVA table

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	966.269	9	107.363	54.593	.000[b]
	Residual	1893.845	963	1.967		
	Total	2860.114	972			

[a] Dependent Variable: commitment
[b] Predictors: (Constant), gender, s21, age, category_3, category_2, s10, s23, s9, s19

After assessing the overall model fit, it is time to look at the individual coefficients shown in Table 7.4. First, we look at the coefficients' *p*-values (**Sig.**) to assess whether these are significant. We find six significant coefficients (*s19, s21, s23, category_3, age*, and *gender*), with *p*-values below 0.05. Note that 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) of the **Unstandardized Coefficients Beta** column in your interpretation. For example, the more satisfied respondents are with Oddjob Airways service delivery (*s19*), the higher their commitment. The unstandardized **B** (or β) is **.012**, which means that when the variable *s19* moves up by one unit (i.e., respondents rate it as 1 higher on average), the dependent variable *commitment* goes up by 0.012 units. The same logic applies to the interpretation of the other significant coefficients. The interpretation *gender* is slightly different as this is a dummy variable with only two categories. The negative coefficient of **–.347** indicates that moving from the base category (female) to the other category (male) will decrease commitment by 0.347 units. That is, males show less commitment to Oddjob Airways than females. We find that the coefficient of *category_3* (*Gold*) is positive (**.010**) and significant. Thus, we conclude that flyers in the *Gold* category show significantly higher levels of commitment than those in the *Blue* category (i.e., the reference category).

Next, we should check the **Standardized Coefficients Beta** column to assess which of the variables have the strongest impact on the dependent variable. This cannot be read from the unstandardized βs! Remember that the standardized β allows us to compare the relative effect of differently measured independent variables by expressing the effect in terms of standard deviation changes from the mean. To interpret the standardized βs, we should look at the highest absolute value first, which is **.245** for *s21* ("Oddjob Airways makes traveling uncomplicated”). This result suggests that to increase their customers’ commitment, Oddjob Airways should focus on making the travelling as uncomplicated as possible. The variable *s23* (“Oddjob Airways offers great value for money”) has the second strongest impact on customer commitment. Note, that while these standardized β coefficients of *age* and *gender* also have a significant bearing on the customers’ commitment, they do not indicate much managerial importance.

In this example, we estimated one model as determined by prior research and management input. However, in other instances, we might have alternative models, which we wish to compare in terms of their fit. In Box 7.4, we describe how to do this by using the relative fit statistic AIC.

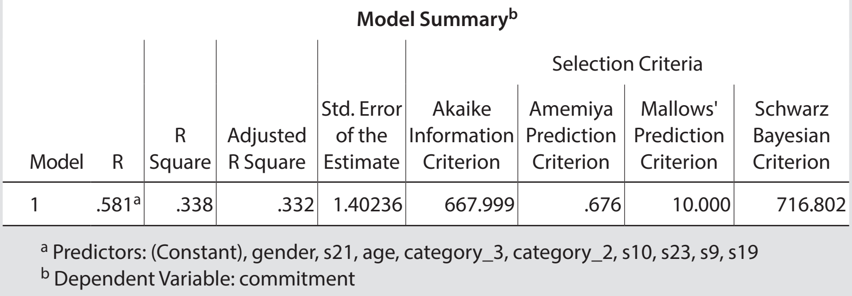
Box 7.4. Model comparison using the AIC and BIC

AIC and BIC are commonly used metrics to compare models with the same dependent variable but different sets of independent variables. Unfortunately, AIC and BIC cannot be accessed by SPSS’s graphical user interface but only via syntax, which is, however, not difficult to do. To show how to request these metrics, simply run the initial regression model as explained earlier. However, in the final step, instead of clicking **OK**, click on **Paste**. This brings up the **IBM Statistics Syntax Editor** in which you will see a command that starts with the word **REGRESSION**. Find the line that starts with/**STATISTICS** and add **SELECTION** to the end of this line. In our case the resulting code should look like this:

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA COLLIN TOL SELECTION
/CRITERIA = PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT commitment
/METHOD = ENTER s9 s10 s19 s21 s23 category_2 category_3 age gender
/SAVE PRED RESID.
```

Then go to ► Run Statistics ► All, which will produce the output shown in ◘ Table 7.9. Under **Akaike Information Criterion**, we find the AIC value, which is 667.999. Similarly, under **Schwarz Bayesian Criterion**, we find the BIC value, which is 716.802. We could now re-run the regression model with a different set-up (e.g., omitting *s9*) and compare the initial model’s AIC and BIC values with those of the new model.

◘ **Table 7.9** Relative measures of fit



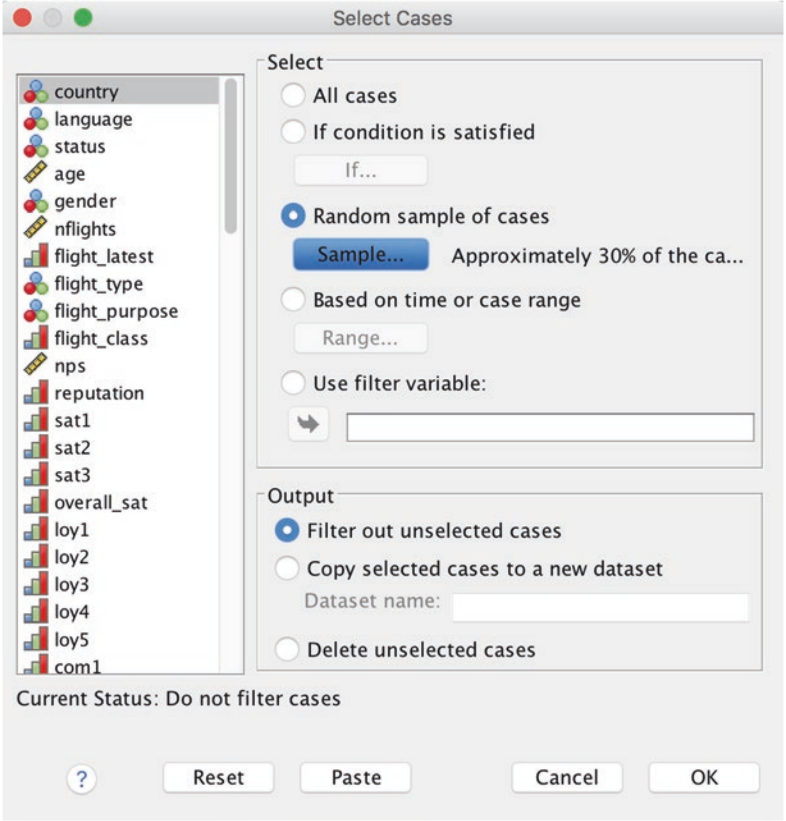
Model Summary[b]

					Selection Criteria			
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion
1	.581[a]	.338	.332	1.40236	667.999	.676	10.000	716.802

[a] Predictors: (Constant), gender, s21, age, category_3, category_2, s10, s23, s9, s19
[b] Dependent Variable: commitment

■ Validate the Regression Model Using SPSS

Next, we need to validate the model. Let's first split-validate our model by randomly selecting 30 % of the cases. Note that we need to make sure that the 30 % of cases left still meet the minimum sample size requirements, else split validation should not be conducted. To do this, go to ► Data ► Select Cases, which will open a dialog box like in ◘ Fig. 7.16. Then click on **Random sample of cases**, followed by **Sample**. Type **30** in the box **Approximately … % of all cases** and click on **Continue**. Before clicking on **OK**, make sure to select **Filter out unselected cases** under **Output**. Now click on **OK**. SPSS will now produce a new variable *filter_$* to the dataset, which takes the value **1** if an observation has been selected, and **0** else. When running an analysis, SPSS now only considers those observations with a value of 1. However, our aim is to compare the regression results from the 30 % of the dataset with the other 70 % of the dataset. Hence, we should go ► Data ► Select Cases and select **All cases**. Next, we need to tell SPSS to run separate regressions for all observations where *filter_$* = 0 versus *filter_$* = 1. This can easily be done by using the split file command, which we can find under ► Data ► Split File. In the dialog box that opens (◘ Fig. 7.17), select the option **Organize output by groups**, move the *filter_$* variable into the **Groups Based on** box, and click on **OK**.

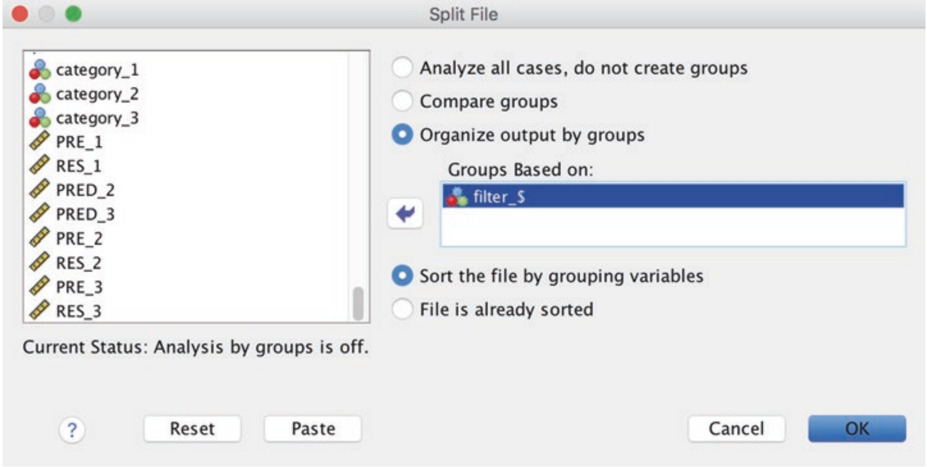

Select Cases

Select
○ All cases
○ If condition is satisfied
● Random sample of cases
Sample... Approximately 30% of the ca...
○ Based on time or case range
○ Use filter variable:

Output
● Filter out unselected cases
○ Copy selected cases to a new dataset
Dataset name:
○ Delete unselected cases

Current Status: Do not filter cases

◘ **Fig. 7.16** The select cases dialog box



Split File

category_1
category_2
category_3
PRE_1
RES_1
PRED_2
PRED_3
PRE_2
RES_2
PRE_3
RES_3

○ Analyze all cases, do not create groups
○ Compare groups
● Organize output by groups

Groups Based on:
filter_$

● Sort the file by grouping variables
○ File is already sorted

Current Status: Analysis by groups is off.

Fig. 7.17 The Split File dialog box

We now have two groups of data, which SPSS analyses separately when re-running the regression analysis. We do not show the model estimation results since the selection of cases is random and therefore will not match the output you obtain! Check if the results obtained from the analysis of the estimation sample (i.e., the 70 %) and validation sample (i.e., the 30 %) are similar. It is likely some effects will change but the general pattern should not change. This means that the signs of the regression coefficients should be the same in the estimation sample and validation sample. Similarly, when sorting the coefficients by size, the ordering should be identical across the samples.

As we have no second dataset available, we cannot re-run the analysis to compare results between datasets. We do, however, have access to other variables such as *country*. If we add this variable, we should check again if the models are similar. If they are, we can conclude that the results are stable.

7.5 Farming with AgriPro (Case Study)

Case Study

AgriPro (http://www.agriprowheat.com) 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 USA locations. In these experiments, the agricultural and economic characteristics, regional adaptation, and yield potential of different varieties of wheat seeds are investigated. In addition, the benefits of the wheat produced, including the milling and baking quality, are examined. If a new variety of wheat seed with superior characteristics is identified, AgriPro produces and markets it throughout the USA and parts of Canada. AgriPro's product is sold to farmers through their distributors, known in the industry as growers. Growers buy wheat seed from AgriPro, grow wheat, harvest the seeds, and sell the seed to local farmers, who plant them in their fields. These growers also provide the farmers, who buy their seeds, with expert local knowledge about management and the environment.

AgriPro sells its products to these growers in several geographically defined markets. These markets are geographically defined, because the 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 AgriPro market, but the company has been performing below the management expectations in it. The heartland region includes the states of Ohio, Indiana, Missouri, Illinois, and Kentucky.

To help AgriPro understand more about farmers in the heartland region, it commissioned a marketing research project involving the farmers in these states. AgriPro, together with a marketing research firm, designed a survey, which included questions regarding what farmers planting wheat find important, how they obtain information on growing and planting wheat, what is important for 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 farm, how much wheat they planted, how old they were, and their level of education.

This survey was mailed to 650 farmers 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, based on the dataset, provide the AgriPro management with advice. This dataset is labeled *Agripro.sav* and is available in the ⤓ Web Appendix (→ ▶ Chap. 7 → Downloads). Note that the dataset contains the variable names and labels matching those in the survey. In the Web Appendix (⤓ Web Appendix → Downloads), we also include the original survey.[15]

© valio84sl/Getty Images/iStock
https://www.guide-market-research.com/app/download/13488671727/SPSS+3rd_Chapter+7_Wheat+farming+survey.pdf?t=1516713162

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

To help you with this task, AgriPro has prepared several questions that it would like to see answered:

1. What do these farmers find important when growing wheat? Please describe the variables *import1* ("Wheat fulfills my rotational needs"), *import2* ("I double crop soybeans"), *import3* ("Planting wheat improves my corn yield"), *import4* ("It helps me break disease and pest cycles"), and *import5* ("It gives me summer cash flow") and interpret.
2. What drives how much wheat these farmers grow (*wheat*)? Agripro management is interested in whether *import1, import2, import3, import4*, and *import5* can explain *wheat*. Please run this regression model and test the assumptions. Can you report on this model to AgriPro's management? Please discuss.
3. Please calculate the AIC and BIC for the model discussed in question 2. Then add the variables *acre* and *age*. Calculate the AIC and BIC again. Which model is better? Should we present the model with or without *acre* and *age* to our client?
4. AgriPro expects that farmers who are more satisfied with their products devote a greater percentage of their total number of acres to wheat (*wheat*). Please test this assumption by using regression analysis. The client has requested that you control for the number of acres of farmland (*acre*), the age of the respondent (*age*), the quality of the seed (*var3*), and the availability of the seed (*var4*), and check the assumptions of the regression analysis. Note that a smaller sample size is available for this analysis, which means the sample size requirement cannot be met. Proceed with the analysis nevertheless. Are all the other assumptions satisfied? If not, is there anything we can do about this, or should we ignore the assumptions if they are not satisfied?
5. Agripro wants you to consider which customers are most loyal to its biggest competitor Pioneer (*loyal5*). Use the number of acres (*acre*), number of acres planted with wheat (*wheat*), and the age of the respondent (*age*). What findings do we obtain? Does this regression model meet the requirements and assumptions?
6. As an AgriPro's consultant, and based on this study's empirical findings, what marketing advice do you have for AgriPro's marketing team? Using bullet points, provide four or five carefully thought through suggestions.

7.6 Review Questions

1. 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, during the first year of their launch. Which independent variables would you use to explain these new products' profitability?

3. Imagine you have to present the findings of a regression model to a client. 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 in the context of regression analysis. Do you agree?
5. Are all regression assumptions equally important? Please discuss.
6. Using standardized βs, we can compare effects between different variables. Can we compare apples and oranges after all? Please discuss.
7. Try adding or deleting variables from the regression model in the Oddjob Airways example and use the adjusted R^2, as well as AIC statistic, to assess if these models are better.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Baum, C. F. (2006). *An introduction to modern econometrics using Stata*. College Station, TX: Stata Press.

Burnham, K. P., & Anderson, D. R. (2013). *Model Selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NJ: Springer.

Cohen, J. (1994). The earth is round (p < .05). *The American Psychologist*, *49*(912), 997–1003.

Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, *70*(1), 1–10.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression, II. *Biometrika*, *38*(1–2), 159–179.

Fabozzi, F. J., Focardi, S. M., Rachev, S. T., & Arshanapalli, B. G. (2014). *The basics of financial econometrics: tools, concepts, and asset management applications*. Hoboken, NJ: John Wiley & Sons.

Field, A. (2013). *Discovering statistics using SPSS* (4th ed.). London: Sage.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*(3), 499–510.

Greene, W. H. (2011). *Econometric analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Hair Jr., J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Boston, MA: Cengage.

Hill, C., Griffiths, W., & Lim, G. C. (2011). *Principles of econometrics* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(3), 305–321.

Mason, C. H., & Perreault Jr., W. D. (1991), Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, *28*(3), 268–280.

Mooi, E. A., & Frambach, R. T. (2009). A stakeholder perspective on buyer–supplier conflict. *Journal of Marketing Channels*, *16*(4), 291–307.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690.

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, *36*(4), 859–866.

Ramsey, J. B. (1969). Test for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B*, *31*(2), 350–371.

Treiman, D. J. (2014). *Quantitative data analysis: Doing social research to test ideas*. Hoboken, NJ: John Wiley & Sons.

VanVoorhis, C. R. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 43–50.

Further Reading

Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, *26*(3), 438–445.

Iacobucci, D. (2008). *Mediation analysis: Quantitative applications in the social sciences*. Thousand Oaks, CA: Sage.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Spiller, S. A., Fitzsimons, G. J., Lynch Jr., J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research*, *50*(2), 277–288.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197–206.