Regression Analysis

7

Keywords

Adjusted R^2 • Akaike information criterion (AIC) • Autocorrelation • Bayes information criterion (BIC) • Binary logistic regression • Breusch-Pagan test • Coefficient of determination • Constant • Collinearity • Cross validation • Disturbance term • Dummy variable • Durbin-Watson test • Error • Estimation sample • η^2 (eta-squared) • F-test • Heteroskedasticity • Interaction effects • Intercept • Moderation analysis • (Multi)collinearity • Multinomial logistic regression • Multiple regression • Nested models • Ordinary least squares • Outlier • Ramsey's RESET test • Residual • R^2 • Robust regression • Simple regression • Split-sample validation • Standard error • Standardized effects • Unstandardized effects • Validation sample • Variance inflation factor (VIF) • White's test

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 Stata.
- How to interpret regression analysis output produced by Stata.

E. Mooi et al., *Market Research*, Springer Texts in Business and Economics,
DOI 10.1007/978-981-10-5218-7_7

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 (Chap. 3). 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 whether independent variables have a significant effect on dependent variables, 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, which 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

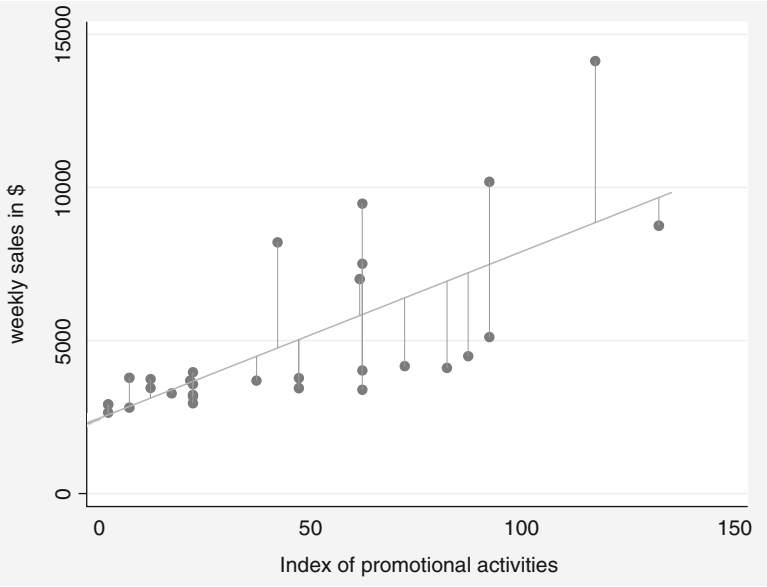

Fig. 7.1 A visual explanation of regression analysis

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 denoted 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. Thus, if the index of promotional activities is zero, we expect the weekly supermarket sales of a specific milk brand to be $2,500. 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 54.59, meaning 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 $54.59. This β_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 β_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 $\widehat{y}$ (pronounced *y*-hat). Thus, the error of each observation is:[1]

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

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 similar to that of simple regression. If we were to have two independent variables, say 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, see 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

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

[2] This only applies to the standardized βs.

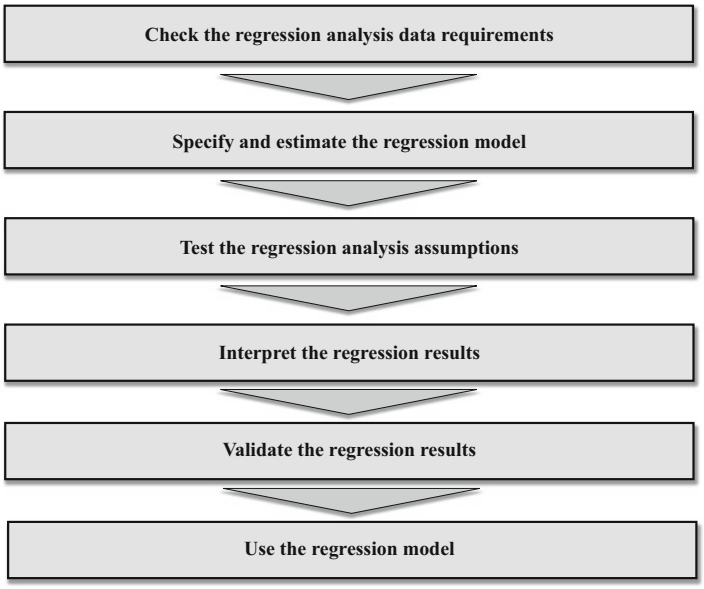

Fig. 7.2 Steps to conduct a regression analysis

how to interpret and validate the regression results. The last step is to use the regression model to, for example, make predictions.

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 6.1), 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 also have to set the power—0.80 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, while: https://stats.idre.ucla.edu/stata/dae/multiple-regression-power-analysis/ discusses how to calculate sample sizes precisely.
- The second approach is by using rules of thumb. These rules are not specific or precise, but are easy to apply. Green (1991) and 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, say less than 10 and certainly less than 15. VanVoorhis and Morgan (2007) add that having at least 30 observations per variable (i.e., $30 \times k$) 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! http://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 with dependent variables measured on a nominal or ordinal scale, see Cameron and Trivedi (2010).

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 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; this means one coefficient will not be estimated. 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**). 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. 2013).[4] Some research now 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 (say 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. If perfect collinearity occurs, drop one of the perfectly overlapping variables. If weaker forms of collinearity occur, you can utilize two 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'

[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.

[4]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.

information, but are uncorrelated. If you use factors, collinearity between the previously collinear variables is no longer an issue.
- The second option is to re-specify the regression model by removing highly correlated variables. Which variables should you remove? If you create a correlation matrix 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—pick the most relevant variable of the two.
- 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.

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 explaining the sales of a particular 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.
- 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.

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 β of the regression model only containing promotions is estimated as 54.59. If we add the variable price (arguably a key variable), the estimated β of promotions drops to 42.27. As can be seen, the difference between the estimated βs in the two models (i.e., with and without price) is 12.32, 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 βs. 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 β_1 estimate. More precisely, the β_1 is likely to be inflated, which means that the estimated value is higher than it should be. Thus, the β_1 itself is biased because we omit x_2!

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.

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. 2008). 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.

Table 7.1 Regression data

Week	Sales	Price	Promotion
1	3,454	1.10	12.04
2	3,966	1.08	22.04
3	2,952	1.08	22.04
4	3,576	1.08	22.04
5	3,692	1.08	21.42
6	3,226	1.08	22.04
7	3,776	1.09	47.04
8	14,134	1.05	117.04
9	5,114	1.10	92.04
10	4,022	1.08	62.04
11	4,492	1.12	87.04
12	10,186	1.02	92.04
13	7,010	1.08	61.42
14	4,162	1.06	72.04
15	3,446	1.13	47.04
16	3,690	1.05	37.04
17	3,742	1.10	12.04
18	7,512	1.08	62.04
19	9,476	1.08	62.04
20	3,178	1.08	22.04
21	2,920	1.12	2.04
22	8,212	1.04	42.04
23	3,272	1.09	17.04
24	2,808	1.11	7.04
25	2,648	1.12	2.04
26	3,786	1.11	7.04
27	2,908	1.12	2.04
28	3,395	1.08	62.04
29	4,106	1.04	82.04
30	8,754	1.02	132.04

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 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 *regression.dta* can be downloaded, but are also included in Table 7.1. (see ↓ Web Appendix → Downloads).

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

$$\widehat{\beta} = \left(x^T x\right)^{-1} . \; x^T y$$

In this equation to solve the $\widehat{\beta}$, 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 by 3 matrix. Next, we multiply this matrix with the untransposed matrix, indicated as x, consisting of the same elements (as a 3 by 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 by 3 matrix $(x^T x)^{-1}$. Next, we calculate $x^T y$, which consists of the 30 by 3 matrix and the vector with the dependent variables' observations (a 1 by 30 matrix). In applied form:[5]

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

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

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

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

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

This last matrix indicates the estimated βs with 30304.05 representing the intercept, -25209.86 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. This shows how the OLS estimator is calculated.

[5]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 $(x^T x)^{-1}$.

As discussed before, each $\widehat{\beta}$ has a standard error, which expresses the uncertainty associated with the estimate. This standard error can be expressed in standard deviations and, as discussed in Chap. 6, with more than 100 degrees of freedom and $\alpha = 0.05$, t-values outside the critical value of ± 1.96 indicate that the estimated effect is *significant* and that the null hypothesis can be rejected. If this the t-value falls within the range of $\pm$ 1.96, the $\widehat{\beta}$ is said to be *insignificant*.

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 heteroskedastic (discussed in Sect. 7.3.3.3), we need to account for this by, for example, using **robust regression** (White 1980).[6] Random-effects estimators allow for estimating a model with correlated errors. There are many more estimators, but these are beyond the scope of this book. Greene (2011) discusses these and other estimation procedures in detail. Cameron and Trivedi (2010) discuss their implementation in Stata.

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, however 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 it becomes we cannot rely 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.

[6] In Stata this can be done by using the `, robust` option.

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 βs) 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. Fig. 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 line corresponds to a x_1^2 transformation. It visually seems that a linear 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 will 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.

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

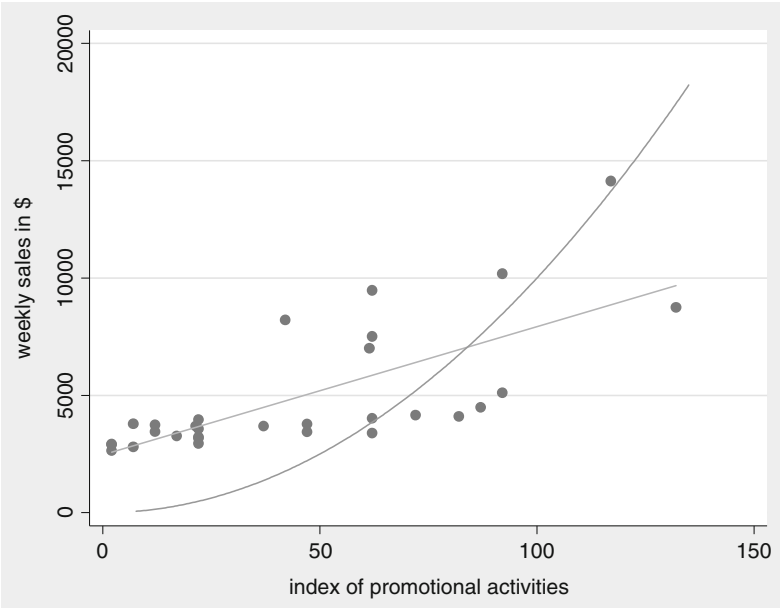

Fig. 7.3 Different relationships between promotional activities and weekly sales

(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(s) has(ve) 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 **heteroskedasticity**. 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 heteroskedasticity. As the dependent variable increases, the error variance also increases. If heteroskedasticity is an issue, the points are typically funnel shaped, displaying more (or less) variance as the independent variable increases (decreases). This funnel shape is typical of heteroskedasticity and indicates that, as a function of the dependent variable, the error variance changes.

We can always try to visualize heteroskedasticity, whose presence is calculated by means of the errors, but it is often very difficult to determine visually whether heteroskedasticity is present. For example, when datasets are large, it is hard to see a funnel shape in the scatterplot. We can formally test for the presence of heteroskedasticity by using the **Breusch-Pagan test** and **White's test**.

The Breusch-Pagan test (1980) is the most frequently used test. This test determines whether the errors' variance depends on the variables in the model by

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

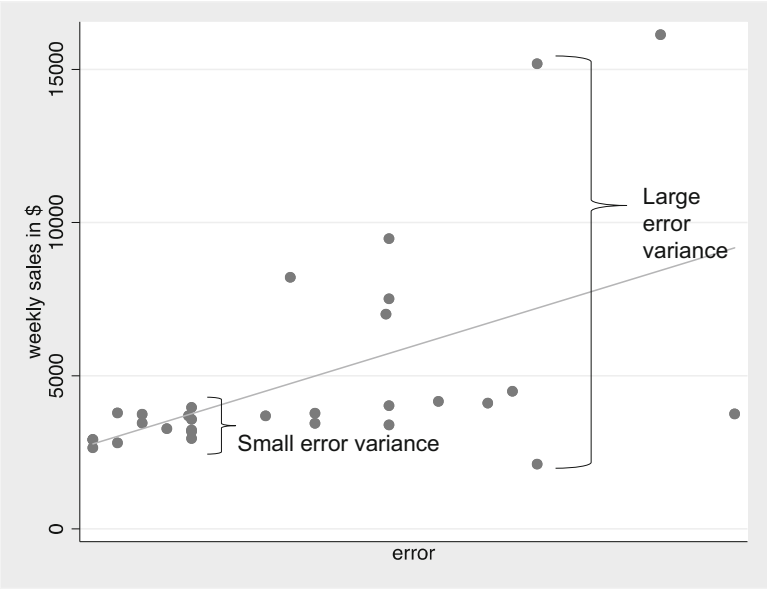

Fig. 7.4 An example of heteroskedasticity

using a separate regression analysis (Greene 2011). This tests the null hypothesis that the errors' variance does not depend on the variables in the regression model.[8] Rejecting this null hypothesis suggests that heteroskedasticity is present. If the Breusch-Pagan test indicates that heteroskedasticity is an issue, *robust regression* remedy for this (see Sect. 7.3.2.2). Note that to illustrate heteroskedasticity, we have used slightly different data than in the other examples.

White's test, as extended by Cameron and Trivedi (1990), is a different test for heteroskedasticity. This test does not test if the error goes up or down, but adopts a more flexible approach whereby errors can first go down, then up (i.e., an hourglass shape), or first go up, then down (diabolo-shaped). Compared to the Breusch-Pagan test, White's test considers more shapes that can indicate heteroskedasticity. This has benefits in that more forms of heteroskedasticity can be detected, but in small samples; however, White's test may not detect heteroskedasticity, even if it is present. It is therefore best to use both tests, as they have slightly different strengths. Generally, the two tests are comparable, but if they are not, it is best to rely on White's test.

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

[8] Specifically, in the mentioned regression model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$, the Breusch-Pagan test determines whether $\widehat{e}^2 = \alpha + \beta_{BP1} x_1 + \beta_{BP2} x_2 + \beta_{BP3} x_3 + e_{BP}$.

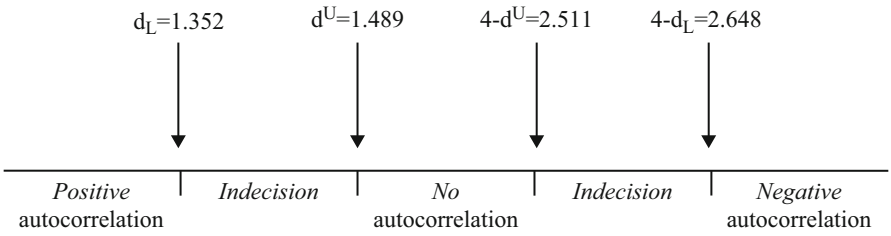

Fig. 7.5 Durbin-Watson test values ($n = 30$, $k = 1$)

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.

First, 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).

Second, 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.

Third, if no systematic pattern of errors occurs, we have no autocorrelation. This absence of autocorrelation is required to estimate standard (OLS) regression models.

Fourth, 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.

1. If the test statistic is lower than the lower critical value ($d < d_L$), we have positive autocorrelation.
2. If the test statistic is higher than 4 minus the lower critical value ($d > 4 - d_L$), we have negative autocorrelation.
3. 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.
4. 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 Cameron and Trivedi (2010) is a useful source of further information.

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 independent and (a) dependent 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 test (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

The model significance is the first aspect that should be determined. While model significance is not an indicator of (close) fit, it makes little sense to discuss model fit if the model itself is not significant. The **F-test** determines the model significance. 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:[9]

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

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. A p-value of the F-test 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.

If we find that the F-test is significant, we can interpret the model fit by using the $\mathbf{R^2}$. 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*. In the scatter plot, we see 30 observations of sales and price (note that we use a small sample size for illustration purposes). The horizontal line (at about $5,000 sales per week) refers to the average sales in 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 squares* (SS_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 $2,500 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_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 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*

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

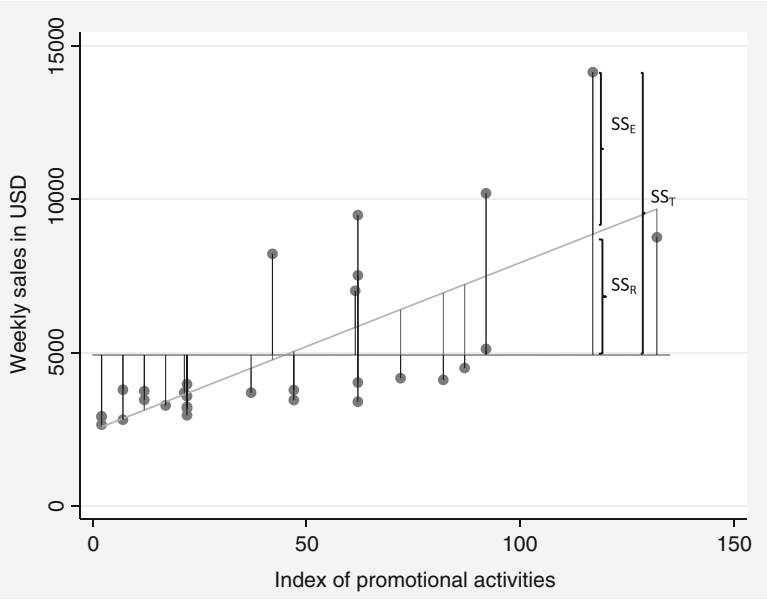

Fig. 7.6 Explanation of the R^2

of squares (SS_E). In more formal terms, we can describe these types of variation as follows:

$$SS_T = SS_R + SS_E$$

This is the same as:

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

Here, n describes the number of observations, y_i is the value of the independent variable for observation i, $\hat{y}_i$ is the predicted value of observation i, and $\bar{y}$ is the mean value of y. As you can see, this description is like the one-way ANOVA we discussed in Chap. 6. A useful regression line should explain a substantial amount of variation (have a high SS_R) relative to the total variation (SS_T):

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

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.

It is difficult to provide rules of thumb regarding what R^2 is appropriate, as this varies from research area to research area. For example, in longitudinal studies, R^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.

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 biased 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:

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

Here, n describes the number of observations and k the number of independent variables (not counting the constant α). This adjusted 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. Furthermore, it is important to note that 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 relative measures 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[10] 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). Smaller values are better and, when comparing models, 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 5.35. 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 the corresponding *p*-value by

[10]The AIC is specifically calculated as AIC $= n \cdot \ln(SS_E/n) + 2 \cdot k$, where n is the number of observations and k the number of independent variables, while the BIC is calculated as BIC $= n \cdot \ln(SS_E/n) + k \cdot \ln(n)$.

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 (54.59) indicates that a one-unit change in (the index of) promotional activities increases sales by 54.59 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 β is 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.

Two further tips: First, only consider significant βs in this respect, as insignificant βs do not (statistically) differ from zero! Second, while the standardized βs are helpful from a practical point of view, standardized βs 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

(continued)

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.[11]

Another way to interpret the size of individual effects is to use the η^2 (pronounced **eta-squared**), which, similar to the R^2, is a measure of the variance accounted for. There are two types of η^2: The model η^2, which is identical to the R^2, and each variable's partial η^2, which describes how much of the total variance is accounted for by that variable. Just like the R^2, the η^2 can only be used to compare variables within a regression model and cannot be used to compare them between regression models. The η^2 relies on different rules of thumb regarding what are small, medium, and large effect sizes. Specifically, an effect of 0.02 is small, 0.15 is medium, and 0.30 and over is large (Cohen 1992).

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 are 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 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

[11] Cohen's (1994) classical article "The Earth is Round ($p < 0.05$)" offers an interesting perspective on significance and effect sizes.

Box 7.2 Moderation

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 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 $x_1 \cdot m$, 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).

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.

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 based on assumed values of the independent variables and their related, 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 price (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 coefficients using regression analysis are 30,304.05 for the intercept, −25,209.86 for price, and 42.27 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:

$$\widehat{y} = 30,304.05 - 25,209.86 \cdot \$1.10 + 42.27 \cdot 50 = \$4,686.70.$$

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 Stata actions.

Table 7.2 Steps involved in carrying out a regression analysis

Theory	Action
Consider the regression analysis data requirements	
Sufficient sample size	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 .10 or lower), use at least $30 \cdot k$ observations per independent variable.
	This can be done easily by calculating the correlation matrix. Note the number of observations (**obs**=. . .) immediately under the correlate command to determine the sample size available for regression.
	`correlate commitment s9 s10 s19 s21 s23 status age gender`
Do the dependent and independent variables show variation?	Calculate the standard deviation of the variables by going to ▶ Statistics ▶ Summaries, tables, and tests ▶ Summary and descriptive statistics ▶ Summary statistics (enter the dependent and independent variables). At the very least, the standard deviation (indicated by **Std. Dev.** in the output) should be greater than 0.
	`summarize commitment s9 s10 s19 s21 s23 i.status age i.gender`
Is the dependent variable interval or ratio scaled?	See Chap. 3 to determine the measurement level.
Is (multi)collinearity present?	The presence of (multi)collinearity can only be assessed after the regression analysis has been conducted (to run a regression model; ▶ Statistics ▶ Linear models and related ▶ Linear regression. Under **Dependent variable** enter the dependent variable and add all the independent variables under the box **Independent variables** and click on **OK**).
	Check the VIF: ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Variance inflation factors. Then click on **Launch** and **OK**. The VIF should be below 10 (although it can be higher, or lower, in some cases; see Sect. 7.3.1.4 for specifics).
	`vif`
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
Estimate the regression model	▶ Statistics ▶ Linear models and related ▶ Linear regression. Under **Dependent variable** enter the dependent variable and add all the independent

(continued)

Table 7.2 (continued)

Theory	Action
	variables under **Independent variables** and click on **OK**.
	`regress commitment s9 s10 s19 s21 s23 i.status age gender`
	Use robust regression when heteroskedasticity is present:
	`regress commitment s9 s10 s19 s21 s23 i.status age gender, robust`
Test the regression analysis assumptions	
Can the regression model be specified linearly?	Consider whether you can write the regression model as: $y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + e$
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. ▶ Graphics ▶ Scatterplot matrix. Then add all the variables and click on **Marker properties** where, under **Symbol**, you can choose **Point** for a clearer matrix. Note that you cannot add variables that start with i. (i.e., categorical variables). Then click on **OK**.
	`graph matrix commitment s9 s10 s19 s21 s23 status age gender, msymbol(point)`
	Conduct Ramsey's RESET test to test for non-linearities. Go to ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Ramsey regression specification-error test for omitted variables. Then click on **Launch** and **OK**.
	`estat ovtest`
Is the expected mean error of the regression model zero?	Choice made on theoretical grounds.
Are the errors constant (homoscedastic)?	Breusch-Pagan test: This can only be checked right after running a regression model. Go to ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Tests for heteroskedasticity (hettest). Then click on **Launch** and then **OK**. Check that the Breusch-Pagan / Cook-Weisberg test for heteroskedasticity is not significant. If it is, you can use robust regression to remedy this.
	`estat hettest`
	White's test: This can only be checked right after running a regression model. Go to ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ information matrix test (imtest). Then click on **Launch** and then **OK**.
	`estat imtest`
Are the errors correlated (autocorrelation)?	This can only be checked after running a regression model and by declaring the time aspect. This means

(continued)

Table 7.2 (continued)

Theory	Action
	you need a variable that indicates how the variables are organized over time. This variable, for example, *week*, should be declared in Stata using the `tsset` command, for example, `tsset week`. Then conduct the Durbin–Watson test. You can select this test by going to ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Durbin-Watson statistic to test for first-order serial correlation. Click on **Launch** and then **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).
	`tsset week`
	`estat dwatson`
Are the errors normally distributed?	This can only be checked after running a regression model. You should first save the errors by going to ▶ Statistics ▶ Postestimation ▶ Predictions ▶ Predictions and their SEs, leverage statistics, distance statistics, etc. Then click on **Launch**. Enter the name of the error variable (we use *error* in this chapter), making sure **Residuals (equation-level scores)** is ticked, and click on **OK**.
	You should calculate the Shapiro-Wilk test to test the normality of the errors. To select the Shapiro-Wilk test, go to Statistics ▶ Summaries, tables, and tests ▶Distributional plots and tests ▶ Shapiro-Wilk normality test. Under **Variables** enter *error* and click on **OK**. Check if the Shapiro-Wilk test under **Prob>z** reports a p-value greater than 0.05.
	To visualize, create a histogram of the errors containing a standard normal curve: ▶ Graphics ▶ Histogram and enter ***error***. Under ▶ Density plots, tick **Add normal-density plot**.
	`predict error, res`
	`swilk error`
	`histogram error, normal`
Interpret the regression model	
Consider the overall model fit	Check the R^2 and significance of the F-value.
Consider the effects of the independent variables separately	Check the (standardized) β. Also check the sign of the β. Consider the significance of the t-value (under `P>|t|` **in the regression table).**
To compare models	Calculate the AIC and BIC ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Information criteria – AIC and BIC. Click on **Launch** and then **OK**.
	Check the AIC and BIC, and ascertain if the simpler model has AIC or BIC values that are at least 2, but

(continued)

Table 7.2 (continued)

Theory	Action
	preferably 10, lower than that of the more complex model.
	`estat ic`
Calculate the standardized effects	Check **Standardized beta coefficients** under the **Reporting** tab of the regression dialog box, which can be found under ► Statistics ► Linear models and related ► Linear regression ► Reporting
	Determine, sequentially, the highest absolute values
Calculate the effect size	Make sure you have used OLS regression (and not robust regression). Then go to ► Statistics ► Postestimation ► Specification, diagnostic, and goodness-of-fit analysis ► Eta-squared and omega-squared effect sizes. Then click on **Launch** and **OK**.
	Interpret each eta squared as the percentage of variance explained (i.e., as that variable's R^2). An effect of individual variables of 0.02 is small, 0.15 is medium, and 0.30 and greater is large.
Validate the model	
Are the results robust?	This can only be done easily using the command window. First create a random variable.
	`set seed 12345`
	`gen validate = runiform() < 0.7`
	Then run the regression model where you first select 70% and then last 30% of the cases. Do this by going to ► Statistics ► Linear models and related ► Linear regression. Then click on **by/if/in** and under **If: (expression)** enter *validate==1*. Then repeat and enter *validate==0*.
	`regress commitment s9 s10 s19 s21 s23 i.status age gender, robust if validate==1`
	`regress commitment s9 s10 s19 s21 s23 i.status age gender, robust if validate==0`
	Compare the model results to ensure they are equal.

7.4 Example

Let's go back to the Oddjob Airways case study and run a regression analysis on the data. Our aim is to explain commitment—the customer's intention to continue the relationship. This variable is formed from three items in the dataset: *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"). Specifically, it is formed by taking the mean of these three variables.[12]

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 (↓ Web Appendix → Downloads) offers several variables that may explain commitment (*commitment*). 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*).

7.4.1 Check the Regression Analysis Data Requirements

Before we start, let's see if we have a sufficient sample size. The easiest way to do this is to correlate all the dependent and independent variables (see Chap. 5) by entering all the variables we intend to include in the regression analysis. To do so go to ▶ Statistics ▶ Summaries, tables, and tests ▶ Summary and descriptive statistics ▶ Correlations and covariances. In the dialog box, enter each variable separately (i.e., *commitment, s9 s10,* etc.) and click on **OK**.

As is indicated by the first line in Table 7.3, the number of observations is 973 (**obs=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 9 independent variables, we satisfy this criterion. Note that if we treat *status* as the categorical variable, which it is, we need to estimate two parameters for *status*—one for the *Silver* and one for the *Gold* category—where *Blue* is the baseline (and cannot be estimated). We thus estimate 10 parameters in total (still satisfying 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. In Table 7.3, we can also examine the pairwise correlations to get an idea of which independent variables relate to the dependent variable and which of them might be collinear.

[12]Using the Stata command `egen commitment=rowmean(com1 com2 com3)`

Table 7.3 Correlation matrix to determine sample size

```
correlate commitment s9 s10 s19 s21 s23 status age gender
(obs=973)

             | commit~t       s9      s10      s19      s21      s23   status      age   gender
-------------+---------------------------------------------------------------------------------
  commitment |   1.0000
          s9 |   0.3857   1.0000
         s10 |   0.3740   0.6318   1.0000
         s19 |   0.4655   0.5478   0.5673   1.0000
         s21 |   0.4951   0.5342   0.5135   0.6079   1.0000
         s23 |   0.4618   0.4528   0.5076   0.5682   0.5367   1.0000
      status |   0.0388   0.0114  -0.0237  -0.0042  -0.0149  -0.1324   1.0000
         age |   0.1552   0.1234   0.1738   0.0965   0.1083   0.1445   0.0187   1.0000
      gender |  -0.0743   0.0203   0.0388   0.0186  -0.0043  -0.0337   0.2128  -0.0047   1.0000
```

Table 7.4 Descriptive statistics to determine variation

```
summarize commitment s9 s10 s19 s21 s23 i.status age i.gender

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  commitment |      1,065    4.163693    1.739216         1          7
          s9 |      1,036    72.23359    20.71326         1        100
         s10 |      1,025    64.53854    21.40811         1        100
         s19 |      1,013    57.21027    21.66066         1        100
         s21 |      1,028    58.96498    22.68369         1        100
-------------+---------------------------------------------------------
         s23 |      1,065    48.93521    22.71068         1        100
             |
      status |
      Silver |      1,065    .2300469    .4210604         0          1
        Gold |      1,065    .1342723    .3411048         0          1
             |
         age |      1,065    50.41972    12.27464        19        101
             |
      gender |
        male |      1,065    .7370892    .4404212         0          1
```

Next we should ascertain if our variables display some variation. This is very easy in Stata when using the `summarize` command (as described in Chap. 5) by going to ▶ Statistics ▶ Summaries, tables, and tests ▶ Summary and descriptive statistics ▶ Summary statistics and by entering the variables in the **Variables** box. This, as shown in Table 7.4, results in output indicating the number of observations per variable and their means and standard deviations, along with their minimum and maximum values. Note that the number of observations in Table 7.3 is **973** and indicates the number of cases in which we fully observe all the variables in the set. However, each variable has a larger number of non-missing observations, which is shown in Table 7.4.

When working with categorical variables, we check whether all observations fall into one category. For example, *status* is coded using thee labels, *Blue*, *Silver*, and *Gold*, representing the values 1, 2, and 3. Having information on two categories makes information on the last category redundant (i.e., knowing that an observation does not fall into the *Silver* or *Gold* category implies it is *Blue*); Stata will therefore only show you one less than the total number of categories. The lowest value (here *Blue*) is

removed by default. Categories can be easier to use when the data are nominal or ordinal. Note that each category is coded 0 when absent and 1 when present. For example, looking at Table 7.4, we can see that **.2300469** or 23% of the respondents fall into the *Silver* category and **.1342723** or 13% into the *Gold* category (implying that 64% fall into the *Blue* category). We also did this with the *gender* variable. You can tell Stata to show categories rather than the actual values by using `i.` in front of the variable's name.

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 commitment. Most researchers would consider this to be interval or ratio scaled, which meets the OLS regression data assumptions.

We should next check for collinearity. While having no collinearity is important, we can only check this assumption after having run a regression analysis. To do so, go to ▶ Statistics ▶ Linear models and related ▶ Linear regression. In the dialog box that follows (Fig. 7.7), enter the dependent variable *commitment* under **Dependent variable** and *s9 s10 s19 s21 s23 i.status age gender* under **Independent variables**. Note that because *status* has multiple levels, using *i.* is necessary to tell you how many observations fall into each category, but be aware that the first level is not shown. Then click on **OK**.

Stata will show the regression output (Table 7.5). However, as the task is to check for collinearity, and not to interpret the regression results, we proceed by going to ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Variance inflation factors. Then click on **Launch** and **OK**.

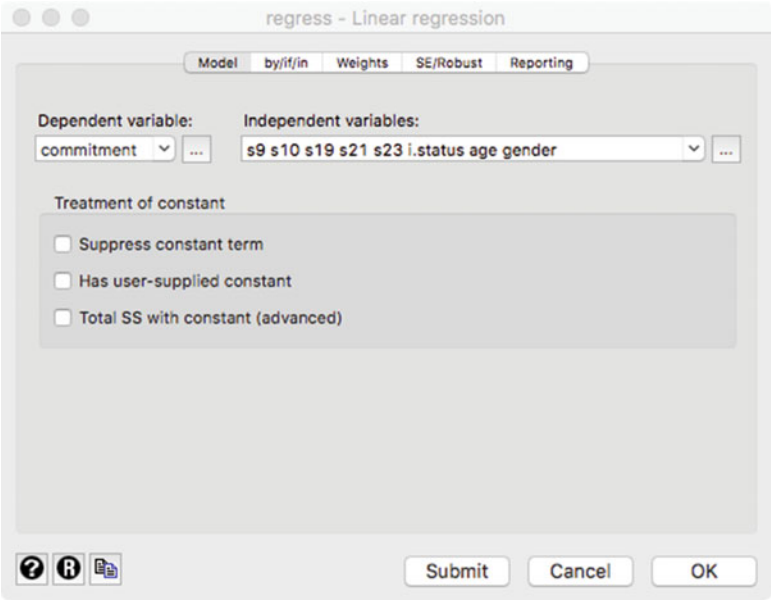

Fig. 7.7 The regression dialog box

Table 7.5 Regression output

```
regress commitment s9 s10 s19 s21 s23 i.status age gender

      Source |       SS           df       MS      Number of obs   =       973
-------------+----------------------------------   F(9, 963)       =     54.59
       Model |  966.268972         9  107.363219   Prob > F        =    0.0000
    Residual |  1893.84455       963  1.96660909   R-squared       =    0.3378
-------------+----------------------------------   Adj R-squared   =    0.3317
       Total |  2860.11352       972  2.94250363   Root MSE        =    1.4024

------------------------------------------------------------------------------
  commitment |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s9 |   .0051594   .0029967     1.72   0.085    -.0007213    .0110401
         s10 |   .0006685   .0029835     0.22   0.823    -.0051865    .0065235
         s19 |   .0122601   .0030111     4.07   0.000      .006351    .0181691
         s21 |   .0186644   .0027365     6.82   0.000     .0132942    .0240345
         s23 |   .0157612   .0026255     6.00   0.000     .0106088    .0209135
             |
      status |
      Silver |    .183402   .1117365     1.64   0.101    -.0358732    .4026771
        Gold |   .4277363   .1377598     3.10   0.002     .1573922    .6980804
             |
         age |   .0102835   .0038561     2.67   0.008     .0027162    .0178509
      gender |  -.3451731   .1050914    -3.28   0.001    -.5514077   -.1389385
       _cons |   1.198751   .2998922     4.00   0.000     .6102336    1.787269
------------------------------------------------------------------------------
```

Table 7.6 Calculation of the variance inflation factors

```
vif

    Variable |       VIF       1/VIF
-------------+----------------------
          s9 |      1.92    0.521005
         s10 |      2.01    0.497696
         s19 |      2.07    0.483267
         s21 |      1.88    0.532030
         s23 |      1.75    0.570861
      status |
           2 |      1.11    0.902297
           3 |      1.12    0.896848
         age |      1.05    0.951353
      gender |      1.06    0.946307
-------------+----------------------
    Mean VIF |      1.55
```

As you can see in Table 7.6, the highest VIF value is **2.07**, which is below 10 and no reason for concern. Note that the individual VIF values are important and not the mean VIF, as individual variables might be problematic even though, on average, collinearity is not a concern. Note that Stata shows the two *status* levels *Silver* and *Gold* as **2** and **3**.

Having met all the described requirements for a regression analysis, our next task is to interpret the regression analysis results. Since we already had to specify a regression model to check the requirement of no collinearity, we know what this regression model will look like!

7.4.2 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*, *status*, *age*, and *gender* as the independent variables. Run the regression analysis again by going to ► Statistics ► Linear models and related ► Linear regression. Having entered the dependent and independent variables in the corresponding boxes, click on the **SE/Robust** tab.

Stata will show you several estimation options (Fig. 7.8). You should maintain the **Default standard errors**, which is identical to **Ordinary least squares (OLS)**. However, when heteroskedasticity is present, use **Robust** standard errors.

Next, click on the **Reporting** tab (Fig. 7.9). Under this tab, you find several options, including reporting the **Standardized beta coefficients**. You can also change the confidence level to 0.90 or 0.99, as discussed in Chap. 6. Under **Set table formats**, you can easily select how you want the regression results to be reported, for example, the number of decimals, US or European notation (1,000.00 vs. 1.000,00), and whether you want to see leading zeros (0.00 vs .00).

Next click on **OK**. This produces the same output as in Table 7.5. Before we interpret the output, let's first consider the assumptions.

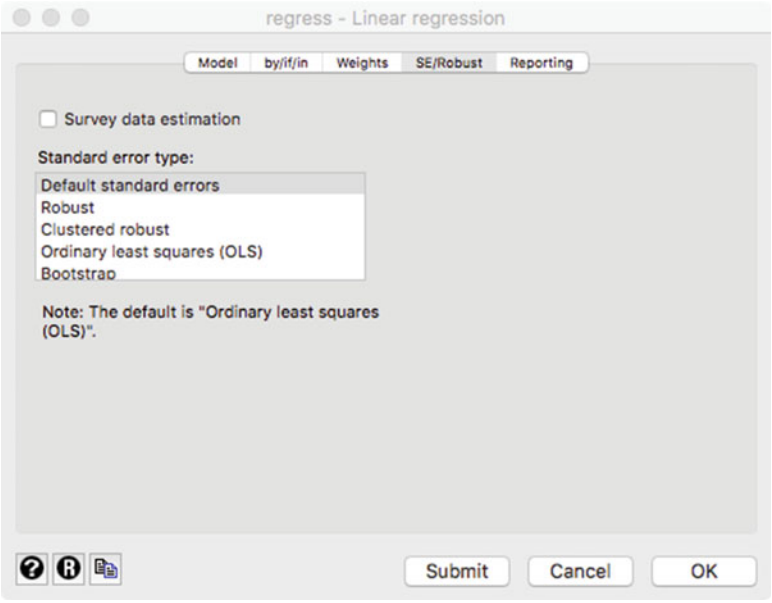

Fig. 7.8 The SE/Robust tab

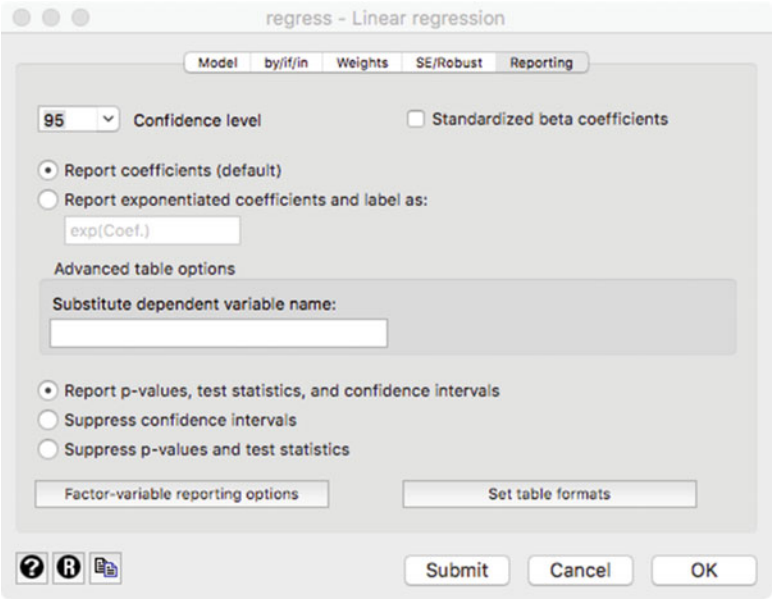

Fig. 7.9 The Reporting tab

7.4.3 Test the Regression Analysis Assumptions

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

$$commitment = \alpha + \beta_1 s9 + \beta_2 s10 + \beta_3 s19 + \beta_4 s21 + \beta_5 s23 + \beta_6 status\_Silver + \beta_7 status\_Gold + \beta_8 age + \beta_9 gender + e$$

Note that because *status* has three levels, two (three minus one) variables are used to estimate the effect of *status*; consequently, we have two βs.

Separately, we also check whether the relationships between the independent and dependent variables are linear. To do this, create a scatterplot matrix of the dependent variable against all the independent variables. This matrix is a combination of all scatterplots (in Chap. 5). To do this, go to ► Graphics ► Scatterplot matrix. Then add all the variables and click on **Marker properties**, where, under **Symbol**, you can choose **Point** for a clearer matrix. Note that you cannot add variables that start with *i*. (i.e., categorical variables) and we therefore just enter *status* (and not *i.status*). Then click on **OK**, after which Stata produces a graph similar to Fig. 7.10. To interpret this graph, look at the first cell, which reads **commitment**. All the scatterplots in the first row (and not the column, as this shows the transpose) show the relationship between the dependent variable and each

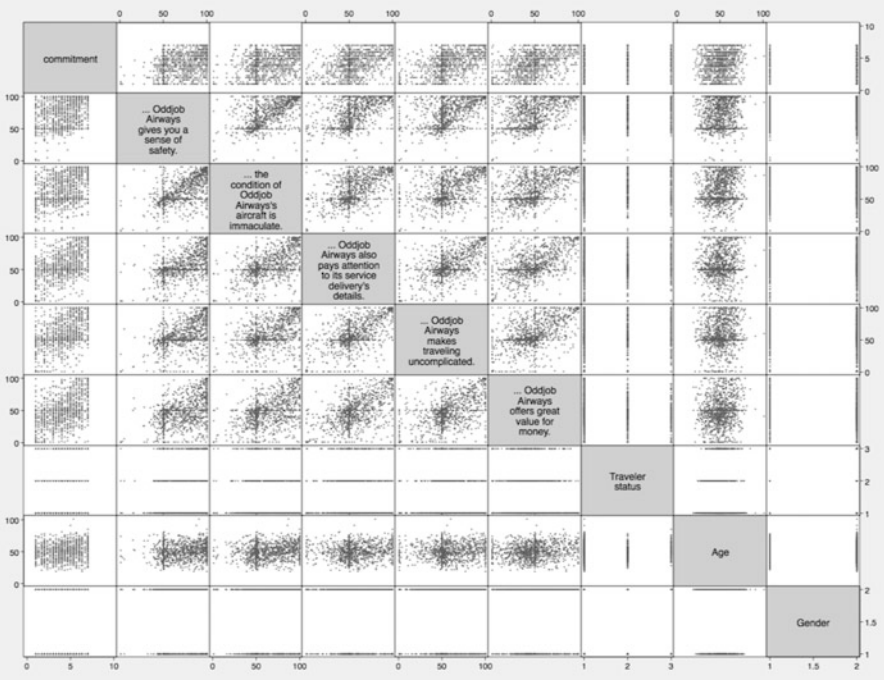

Fig. 7.10 A scatterplot matrix of the dependent variable against all the independent variables

independent variable. The large number of dots makes it difficult to see whether the relationships are linear. However, linearity is also not clearly rejected. Note that the cells **Traveler status** and **Gender** show three (two) distinct bands. This is because these variables take on three distinct values (*Blue*, *Silver*, and *Gold*) for *status* and two (*female* and *male*) for *gender*. When an independent variable has a small number of categories, linearity is not important, although you can still see the form that the relationship might take.

We can also test for the presence of nonlinear relationships between the independent and dependent variable by means of Ramsey's RESET test. Go to ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Ramsey regression specification-error test for omitted variables. Then click on **Launch** and **OK**.

The results in Table 7.7 of this test under **Prob > F** suggest no non-linearities are present, as the *p*-value (**0.3270**) is greater than 0.05. Bear in mind, however, that this test does not consider all forms of non-linearities.

To check the second assumption, we should assess whether the regression model's expected mean error is zero. Remember, this choice is made on theoretical grounds and 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

Table 7.7 Ramsey's RESET test

```
estat ovtest

Ramsey RESET test using powers of the fitted values of commitment
       Ho:  model has no omitted variables
                  F(3, 960) =      1.15
                    Prob > F =      0.3270
```

Table 7.8 Breusch-Pagan test for heteroskedasticity

```
estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of commitment

         chi2(1)      =     8.59
         Prob > chi2  =   0.0034
```

explaining commitment. This makes it highly likely that the regression model's expected mean error is zero.

The third assumption is that of homoscedasticity. To test for this, use the Breusch-Pagan test and go to ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Tests for heteroskedasticity. Then click on **Launch** and **OK**.

The output in Table 7.8 shows the results of the Breusch-Pagan / Cook-Weisberg test for heteroskedasticity. With a *p*-value (**Prob** > **chi2**) of **0.0034**, we should reject the null hypothesis that the error variance is constant, thus suggesting that the error variance is not constant.

To test for heteroskedasticity by means of White's test, go to ▶ Statistics ▶ Postestimation ▶ Specification, diagnostic, and goodness-of-fit analysis ▶ Information matrix test (imtest). Then click on **Launch** and then **OK**.

As you can see in Table 7.9, the output consists of four tests. Under **Heteroskedasticity**, the first element is the most important, as it gives an indication of whether heteroskedasticity is present. The null hypothesis is that there is no heteroskedasticity. In Table 7.9, this null hypothesis is rejected and the findings suggest that heteroskedasticity is present. Both the Breusch-Pagan and White's test are in agreement. Consequently, we should use a robust estimator; this option is shown in Fig. 7.8. as **Robust**. Please note that because we now use a robust estimator, Stata no longer shows the adjusted R^2 and we can only use the AIC and BIC to compare the models.

If we had data with a time component, we would also perform the Durbin-Watson test to check for potential autocorrelation (fourth assumption). This

Table 7.9 White's test for heteroskedasticity

```
estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df       p
---------------------+-----------------------------
  Heteroskedasticity |      78.04     50    0.0068
            Skewness |      30.73      9    0.0003
            Kurtosis |       6.70      1    0.0097
---------------------+-----------------------------
               Total |     115.47     60    0.0000
---------------------------------------------------
```

requires us to first specify a time component, which is absent in the dataset; however, if we had access to a time variable, say *week*, we could time-set the data by using `tsset week`. We can then use the command `estat dwatson` to calculate the Durbin-Watson d-statistic and check whether autocorrelation is present. However, since the data do not include any time component, we should not conduct this test.

Lastly, we should explore how the errors are distributed. We first need to save the errors to do so by going to ▶ Statistics ▶ Postestimation ▶ Predictions ▶ Predictions and their SEs, leverage statistics, distance statistics, etc. Then click on **Launch**. In the dialog box that opens (Fig. 7.11), enter *error* under **New variable name** and tick **Residuals (equation-level scores)**, which is similar to what we discussed at the beginning of this chapter. There are also several other types of variables that Stata can save. The first option, **Linear prediction (xb)**, saves the predicted values. The other options are more advanced and discussed in detail in the Stata Manual (StataCorp 2015).

Next, click on **OK**. Stata now saves the errors so that they can be used to test and visualize whether they are normally distributed. To test for normality, you should run the Shapiro-Wilk test (Chap. 6) by going go to ▶ Statistics ▶ Summaries, tables, and tests ▶ Distributional plots and tests ▶ Shapiro-Wilk normality test. In the dialog box that opens (Fig. 7.12), enter *error* under **Variables** and click on **OK**. The output in Table 7.10 indicates a *p*-value (**Prob > z**) of **0.08348**, suggesting that the errors are approximately normally distributed. Hence, we can interpret the regression parameters' significance by using *t*-tests.

We also create a histogram of the errors comprising a standard normal curve. To do so, go to ▶ Graphics ▶ Histogram and enter *error* under **Variable**. Click on the **Density plots** tab and tick **Add normal-density plot**. The chart in Fig. 7.13 also suggests that our data are normally distributed, as the bars indicating the frequency of the errors generally follow a normal curve.

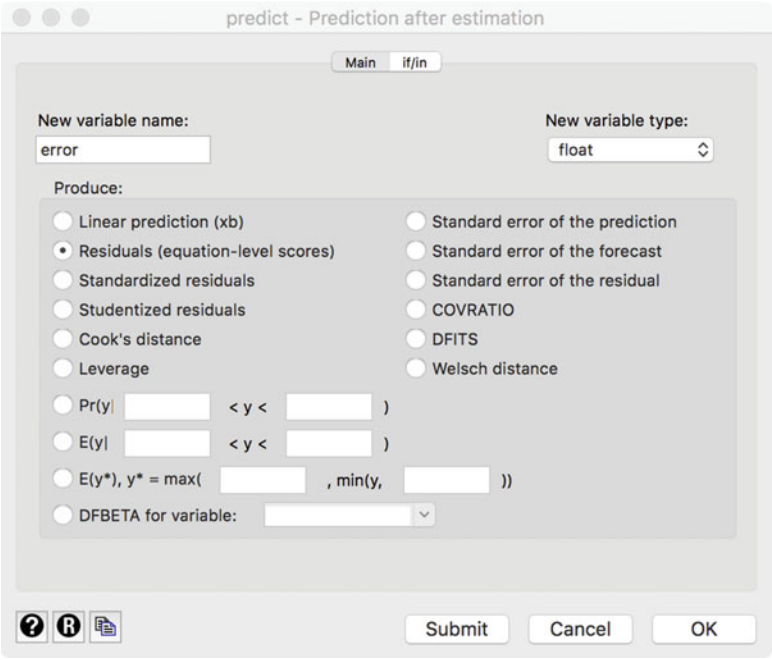

Fig. 7.11 Saving the predicted errors

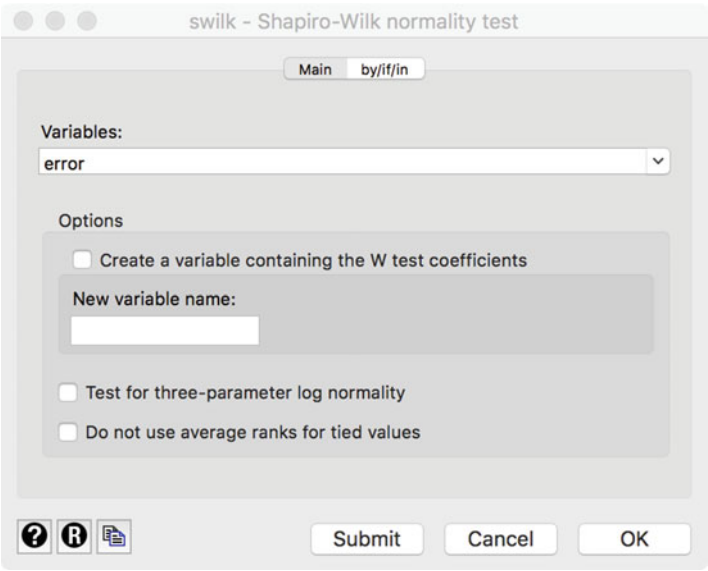

Fig. 7.12 Test for the errors' normality

Table 7.10 The Shapiro-Wilk test for normality

```
swilk error

                    Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V          z       Prob>z
-------------+--------------------------------------------------------
       error |        973    0.99716      1.748      1.382    0.08348
```

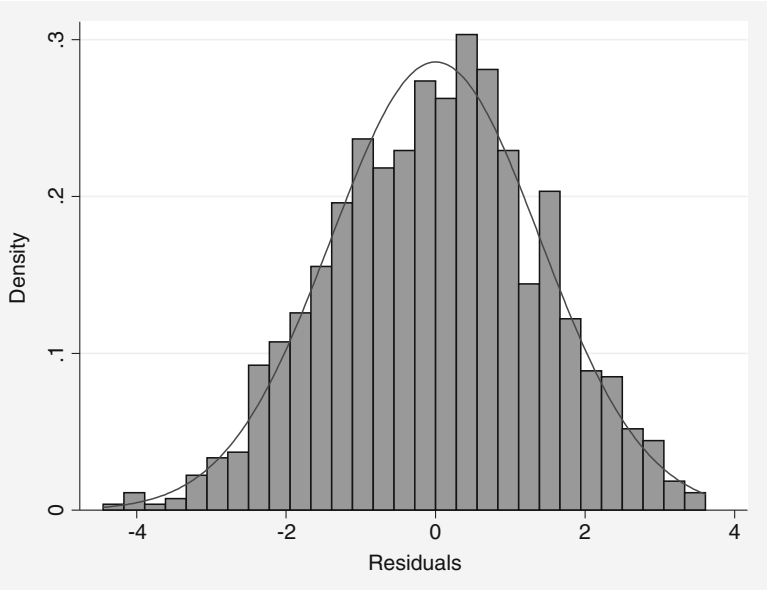

Fig. 7.13 A histogram and normal curve to visualize the error distribution

7.4.4 Interpret the Regression Results

Although we have already conducted a regression analysis to test the assumptions, let's run the analysis again, but this time with robust standard errors because we found evidence of heteroskedasticity. To run the regression, go to ▶ Statistics ▶ Linear models and related ▶ Linear regression. Under **Dependent variable**, enter the dependent variable *commitment* and add all the independent variables *s9 s10 s19 s21 s23 i.status age gender* under **Independent variables**. Then click on **SE/Robust**, select **Robust**, followed by **OK**. Table 7.11 presents the regression analysis results.

Table 7.11 has of two parts; on top, you find the overall model information followed by information on the individual parameters (separated by -----). In the section on the overall model, we first see that the number of observations is 973. Next is the *F*-test, whose *p*-value of **0.000** (less than 0.05) suggests a significant model.[13] Further down, we find that the model yields an R^2 value of **0.3378**, which seems satisfactory and is above the value of 0.30 that is common for cross-sectional research.

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

Table 7.11 Regression output

```
regress commitment s9 s10 s19 s21 s23 i.status age gender, robust

Linear regression                               Number of obs     =        973
                                                F(9, 963)         =      80.20
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3378
                                                Root MSE          =     1.4024

------------------------------------------------------------------------------
             |               Robust
  commitment |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s9 |   .0051594   .0031804     1.62   0.105    -.0010819    .0114007
         s10 |   .0006685   .0032532     0.21   0.837    -.0057157    .0070527
         s19 |   .0122601   .0032491     3.77   0.000      .005884    .0186361
         s21 |   .0186644    .002807     6.65   0.000     .0131558     .024173
         s23 |   .0157612   .0027849     5.66   0.000     .0102961    .0212263
             |
      status |
     Silver  |    .183402   .1152451     1.59   0.112    -.0427585    .4095624
       Gold  |   .4277363   .1333499     3.21   0.001     .1660463    .6894263
             |
         age |   .0102835    .003807     2.70   0.007     .0028125    .0177546
      gender |  -.3451731   .1029958    -3.35   0.001    -.5472952    -.143051
       _cons |   1.198751    .289718     4.14   0.000     .6301998    1.767303
------------------------------------------------------------------------------
```

In the section on the individual parameters, we find, from left to right, information on the included variables (with the dependent *commitment* listed on top), the coefficients, the robust standard errors, the *t*-values and associated *p*-values (indicated as $\mathbf{P > |t|}$), and the confidence intervals. First, you should look at the individual coefficients. For example, for *s19*, we get an effect of **0.0122601**, suggesting that when variable *s19* moves up by one unit, the dependent variable *commitment* goes up by **.0122601** units. Under $\mathbf{P > |t|}$, we find that the regression coefficient of *s19* is significant, as the *p*-value is smaller **0.105** is greater than 0.05. Conversely, *s9* has no significant effect on commitment, as the corresponding *p*-value of **0.105** is greater than 0.05. Further analyzing the output, we find that the variables *s21*, *s23*, *status Gold*, *age*, and *gender* have significant effects. Note that of all the *status* variables, only the coefficient of the tier *status Gold* is significant, whereas the coefficient of *status Silver* is not. A particular issue with these categorical variables is that if you change the base category to, for example, *Silver*, neither *Gold* nor *Blue* will be significant. Always interpret significant findings of categorical variables in relation to their base category. That is, you can claim that *Gold* status travelers have a significantly higher commitment than those of a *Blue*

status.[14] The coefficient of *gender* is significant and negative. Because the variable *gender* is scaled 0 (female) to 1 (male), this implies that males (the higher value) show less commitment to Oddjob Airways than females. Note that because the *gender* variable is measured binary (i.e., it is a *dummy variable*; see Chap. 5), it is always relative to the other gender and therefore always significant. Only when there are 3 or more categories does the interpretation issue, which we saw regarding *status*, occur. Specifically, on average, a male customer shows −**.3451731** units less commitment.

In this example, we estimate one model as determined by prior research and the company management input. However, in other instances, we might have alternative models, which we wish to compare in terms of their fit. In Box 7.3, we describe how to do this by using the relative fit statistics AIC and BIC.

Next, we should check the standardized coefficients and effect sizes to get an idea of which variables are most important. This cannot be read from the *t*-values or *p*-values! To calculate the standardized β coefficients, return to ► Statistics ► Linear models and related ► Linear regression. In the **Reporting** tab, check the **Standardized beta coefficients** and click on **OK**. This will produce the output in Table 7.13. To interpret the standardized β coefficents, look at the largest absolute number, which is **.4277363** for the variable *status Gold*. The second highest value relates to *gender* (−**.3451731**) and is binary. While the third-highest is *s21* ("Oddjob Airways makes traveling uncomplicated"). These variables contribute the most in this order.[15] Note, however, that *gender* is not a variable that marketing

Box 7.3 Model Comparison Using AIC and BIC

When comparing different models with the same dependent variable (e.g., *commitment*), but with different independent variables, we can compare the models' adequacy by means of the AIC and BIC statistics. To do this, go to Statistics ► Postestimation ► Specification, diagnostic, and goodness-of-fit analysis ► Information criteria – AIC and BIC. Click on **Launch** and then **OK**. Stata will then show the output as in Table 7.12. The AIC and BIC are respectively listed as 3429.253 and 3478.057. Remember that the AIC and BIC can be used to compare different models. For example, we can drop *age* and *gender* from the previous model and calculate the AIC and BIC again. Although we do not show this output, the resultant AIC and BIC would then respectively be 3443.283 and 3482.326, which are higher, indicating worse fit and suggesting that our original specification is better.

(continued)

[14]Note that it is possible to show all categories for regression tables by typing `set showbaselevels on`. This can be made permanent by typing `set showbaselevels on, permanent`.

[15]Note that while the constant has the highest value (1.19), this is not a coefficient and should not be interpreted as an effect size.

Box 7.3 (continued)

Table 7.12 Relative measures of fit

```
estat ic

Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+----------------------------------------------------------------
           . |        973 -1905.187  -1704.627      10    3429.253   3478.057
------------------------------------------------------------------------------
```

Table 7.13 Regression output

```
regress commitment s9 s10 s19 s21 s23 i.status age gender, vce(robust) beta

Linear regression                               Number of obs     =        973
                                                F(9, 963)         =      80.20
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3378
                                                Root MSE          =     1.4024

------------------------------------------------------------------------------
             |               Robust
  commitment |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          s9 |   .0051594   .0031804     1.62   0.105                .0625475
         s10 |   .0006685   .0032532     0.21   0.837                 .008328
         s19 |   .0122601   .0032491     3.77   0.000                .1535821
         s21 |   .0186644    .002807     6.65   0.000                .2451992
         s23 |   .0157612   .0027849     5.66   0.000                .2083433
             |
      status |
     Silver  |    .183402   .1152451     1.59   0.112                .0453108
       Gold  |   .4277363   .1333499     3.21   0.001                .0859729
             |
         age |   .0102835    .003807     2.70   0.007                .0716953
      gender |  -.3451731   .1029958    -3.35   0.001               -.0885363
       _cons |   1.198751    .289718     4.14   0.000                       .
------------------------------------------------------------------------------
```

managers can change and the number of people flying determines the *status*. Making flying with Oddjob airways less complicated may be something marketing managers can influence.

To obtain a better understanding of the effect sizes, we can calculate the η^2. Effect sizes can only be calculated when OLS regression and not robust regression is used. Therefore, run the regression model without the `, robust` option and go to ► Statistics ► Postestimation ► Specification, diagnostic, and goodness-of-fit analysis ► Eta-squared and omega-squared effect sizes. Then click on **Launch** and **OK**. Stata will calculate the effect sizes, as shown in Table 7.14. These effect

Table 7.14 Effect sizes

```
estat esize

Effect sizes for linear models

------------------------------------------------------------------------
             Source |   Eta-Squared     df     [95% Conf. Interval]
--------------------+---------------------------------------------------
              Model |   .3378429         9      .286636    .3754686
                    |
                 s9 |   .0030688         1            .    .0138597
                s10 |   .0000521         1            .    .0041106
                s19 |   .0169237         1     .0045586    .0364154
                s21 |   .0460813         1     .0236554    .0743234
                s23 |   .0360722         1      .016498    .0619021
             status |   .0107431         2     .0010116    .0259672
                age |   .0073311         1     .0005013     .021729
             gender |   .0110784         1     .0017974    .0277598
------------------------------------------------------------------------
```

sizes can be interpreted as the R^2, but for each individual variable in that specific model. First, the overall η^2 of **.3378429** is identical to the model R^2 as shown in Table 7.11. We should consider the largest value of the individual variables (*s21*) as the most important variable, because it contributes the most to the explained variance (4.6% or, specifically, **.0460813**). Although this is the largest value, Cohen's (1992) rules of thumb suggest this is a small effect size.[16]

7.4.5 Validate the Regression Results

Next, we need to validate the model. Let's first split-validate our model. This can only be done by means of easy instructions in the command window. First, we need to create a variable that helps us select two samples. A uniform distribution is very useful for this purpose, which we can make easily by typing `set seed 12345` in the command window (press enter), followed by `gen validate=runiform() < 0.7`. The first command tells Stata to use random numbers, but since we fix the "seed," we can replicate these numbers later.[17] The second part makes a new variable called *validate*, which has the values zero and one. We can use this variable to help Stata select a random 70% and 30% of cases. This requires us to run the regression model again and selecting the first 70% and the last 30% of the cases. Let's first estimate our model over the 70% of cases. Do this by going to ► Statistics ► Linear models and related ► Linear regression. Then click on **by/if/**

[16]Please note that only Stata 13 or above feature built-in routines to calculate η^2.

[17]The seed specifies the initial value of the random-number generating process such that it can be replicated later.

Table 7.15 Assessing robustness

```
regress commitment s9 s10 s19 s21 s23 i.status age gender if validate==1, vce(robust)

Linear regression                               Number of obs     =        687
                                                F(9, 677)         =      53.40
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3330
                                                Root MSE          =     1.3969

------------------------------------------------------------------------------
             |               Robust
  commitment |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s9 |   .0037881    .003764     1.01   0.315    -.0036024    .0111786
         s10 |   .0051292   .0038798     1.32   0.187    -.0024886     .012747
         s19 |   .0095681   .0037821     2.53   0.012     .0021421     .016994
         s21 |   .0176205   .0033582     5.25   0.000     .0110269    .0242142
         s23 |   .0155433   .0033412     4.65   0.000      .008983    .0221036
             |
      status |
        Blue |          0  (base)
      Silver |   .1400931   .1376231     1.02   0.309    -.1301264    .4103126
        Gold |   .3958519   .1634049     2.42   0.016     .0750106    .7166931
             |
         age |   .0099883   .0044713     2.23   0.026      .001209    .0187676
      gender |  -.2638042    .124435    -2.12   0.034    -.5081291   -.0194794
       _cons |   1.177727   .3408402     3.46   0.001     .5084958    1.846958
------------------------------------------------------------------------------

regress commitment s9 s10 s19 s21 s23 i.status age gender if validate==0, vce(robust)

Linear regression                               Number of obs     =        286
                                                F(9, 276)         =      34.58
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3694
                                                Root MSE          =     1.4105

------------------------------------------------------------------------------
             |               Robust
  commitment |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s9 |   .0070651   .0060363     1.17   0.243     -.004818    .0189482
         s10 |  -.0091891   .0058937    -1.56   0.120    -.0207915    .0024132
         s19 |   .0186433   .0063824     2.92   0.004     .0060788    .0312078
         s21 |   .0223386   .0048513     4.60   0.000     .0127884    .0318888
         s23 |   .0157948    .005125     3.08   0.002     .0057057    .0258839
             |
      status |
        Blue |          0  (base)
      Silver |   .2740726   .2103388     1.30   0.194    -.1399995    .6881447
        Gold |   .5126804   .2389762     2.15   0.033     .0422327    .9831281
             |
         age |   .0099737   .0071877     1.39   0.166    -.0041761    .0241234
      gender |  -.5085486   .1843601    -2.76   0.006    -.8714793   -.1456179
       _cons |   1.245528   .5731065     2.17   0.031     .1173124    2.373743
------------------------------------------------------------------------------
```

in and under **IF:(expression)** enter *validate==1* and click on **OK**. Stata will now estimate the regression model using 70% of the observations (i.e., the estimation sample). When repeating the process for the remaining 30%, enter *validate==0* under **IF:(expression)** (i.e., the validation sample). Table 7.15 shows the results of these two model estimations. As we can see, the models are quite similar (also when compared to the original model), but the effect of *age* is not significant in the second

model, although the coefficient is very similar. This suggests that the effects are robust.

As we have no second dataset available, we cannot re-run the analysis to compare. We do, however, have access to other variables such as *country*. If we add this variable, all the variables that were significant at $p < 0.05$ remain significant, while *country* is not significant. Based on this result, we can conclude that the results are stable.

7.5 Farming with AgriPro (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.dta* 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.[18] 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. 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 of 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*), the age of the respondent (*age*), and this person's education. Use the `i.` operator for education to gain an understanding of the group differences. 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.

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

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 and BIC statistics, to assess if these models are better.

7.7 Further Readings

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis. A global perspective* (7th ed.). Upper Saddle River: Pearson Prentice Hall.

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

Nielsen at http://www.nielsen.com

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

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

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

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

This is a very good introduction to single and multiple regression. It discusses categorical independent variables in great detail while using Stata.

http://www.ats.ucla.edu/stat/stata/topics/regression.htm

This is an excellent and detailed website dealing with more advanced regression topics in Stata.

References

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

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

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies, 47*(1), 239–253.

Cameron, A.C. & Trivedi, P.K. (1990). *The information matrix test and its implied alternative hypotheses*. (Working Papers from California Davis – Institute of Governmental Affairs, pp. 1–33).

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics using stata* (Revised ed.). College Station: Stata Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

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: Wiley.

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: Prentice Hall.

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis*. Upper Saddle River: Pearson.

Hill, C., Griffiths, W., & Lim, G. C. (2008). *Principles of econometrics* (3rd ed.). Hoboken: Wiley.

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, W. D., Jr. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research, 28*, 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 and Quantity, 41*(5), 673–690.

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.

Sin, C., & White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics, 71*(1–2), 207–225.

StataCorp. (2015). *Stata 14 base reference manual*. College Station: Stata Press.

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

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

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society, 48*(4), 817–838.