Regression Analysis: Assumptions and Diagnostics

In: The SAGE Handbook of Regression Analysis and Causal Inference

By: Bart Meuleman, Geert Loosveldt & Viktor Emonds
Edited by: Henning Best & Christof Wolf
Pub. Date: 2013
Access Date: August 14, 2020
Publishing Company: SAGE Publications Ltd
City: London
Print ISBN: 9781446252444
Online ISBN: 9781446288146
DOI: https://dx.doi.org/10.4135/9781446288146
Print pages: 83-110

Regression Analysis: Assumptions and Diagnostics

BartMeuleman

GeertLoosveldt

ViktorEmonds

Introduction

As shown in the previous chapter, ordinary least squares (OLS) regression links the values of dependent variable Y_i(i = 1, 2,…, n) to the values of a set of independent variables X_{ik} by means of a linear function and an error term ε_i:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i = \sum_k \beta_k X_{ik} + \epsilon_i, \qquad (5.1)$$

where k ranges from 0 to p-1. This model thus contains p regression parameters (namely effects of p −1 predictors and one intercept: X' equals 1 for all cases). The linear function is called the linear predictor or the structural part of the model, while the error term is the random or stochastic component of the model. In general, regression analysis can be used for two purposes: (1) to describe the data structure or (2) to make inferences about the population parameters of the regression model.

Regression analysis can only perform these functions optimally, however, if certain conditions are fulfilled. This chapter deals with the assumptions on which the OLS regression model as described above is built.[1] The exact number of assumptions (and the way in which they are categorized) varies considerably across regression textbooks. In this account, we will limit ourselves to six assumptions we believe to be the most important and widely cited ones. A first group of four classical assumptions follows from the statistical theory underlying regression analysis. The relationships between dependent and independent variables are assumed to be linear (1). Furthermore, it is required that error terms are homoscedastic (2), independent (3) and normally distributed (4). If these assumptions are met, the Gauss–Markov theorem guarantees that the OLS coefficients are the best linear unbiased estimators (BLUE). Here, ‘best’ means that these estimators have the smallest mean squared error. Violations of these four classical assumptions, however, are not the only factors that can hamper regression analysis. In addition, accurate regression analysis requires that predictors are not multicollinear (5) and that influential observations are absent (6).

In the discussion of the six assumptions, we follow a similar structure. First, we indicate the role the respective assumption plays in the regression machinery, and how violations can affect conclusions. Subsequently, we discuss how violations of the assumptions can be diagnosed, and how possible violations can be remedied.

To make the diagnostics and remedies accessible for applied researchers, we provide illustrations using the

fifth round of the European Social Survey (this data set can be downloaded from http://ess.nsd.uib.no/ess/round5/download.html). For the Belgian subsample, we examine a linear regression model with trust in the police (trstplc, ranging from 0 (no trust at all) to 10 (complete trust)) as dependent variable. Predictors in the model are age in years (age), gender (female), number of years of full-time education completed (eduyrs), subjective income in four categories (hincfel – higher scores indicate a lower subjective income), and an evaluation (from 0 to 10) of how successful the police are at preventing criminality in the respondent's country (plcpvcr). The results of this regression (see the output in Table 5.1) show that older people and those with a lower subjective income tend to have a lower trust in the police. Conversely, higher trust levels are found among the higher educated and persons who positively evaluate the ability of the police to prevent crime.

Table 5.1 Regression output

Variable	Parameter	Standard error
Intercept	3.819***	(0.287)
Age	–0.007**	(0.003)
Female	0.003	(0.092)
Hincfel	–0.268***	(0.056)
Eduyrs	0.052***	(0.013)
Plcpvcr	0.434***	(0.025)
Adjusted R^2	.177	

*$p < 0.05$

**$p < 0.01$

***$p < 0.001$

The online appendix to this chapter provides Stata code that can be used to reproduce these examples.

Assumption 1: Linearity

What Is It?

The assumption of linearity is essential for using regression analysis as a descriptive tool. When we use the OLS regression model to describe the relationship between the dependent variable Y and a set of p — 1 independent variables X, we assume that Y is a linear function of X. The basic regression model is fully or completely linear, meaning that the model is linear in both its parameters and its variables. The qualifying characteristic of a model that is linear in the parameters is that a unit change in any parameter value leads to the same change in the dependent variable whatever the values of the parameters. So linearity is a

characteristic of the parameters of the model (Krzanowski, 1998). The basic regression model is also linear in the variables: a unit change in one of the variables produces a constant change in the dependent variables whatever the value of the variable.

The linearity assumption can also be looked at from the perspective of the error terms. There is a perfect linear relationship between Y and X when all error terms in the model equal zero. In this situation, all observations characterized by their coordinates ($x1$, $x2$, …, x_k, y) lie on a hyperplane in a (k + 1)-dimensional space. In bivariate regression, for example, perfect linearity means that all observations are positioned on a straight line in a two-dimensional space. When there are two independent variables, all observations should be situated on a two-dimensional surface in a three-dimensional space. In realistic research settings, however, perfect linear relationships do not occur, and the errors represent the deviations from the perfect linear model. For each covariate pattern (i.e. a specific combination of X-values), there will be several error terms ε_i which can be considered as the difference between the observed Y_i-values and the predicted value for each unit based on the specific values of X_{ik} in the structural part of the model. Now linearity implies that, for each covariate pattern, positive and negative errors balance each other out. Conditional on the X's, the expected value of the errors should equal zero. Or in other words, the linearity assumption stipulates that the conditional means of the dependent variable Y (i.e. given the values of X) equal the predicted values of Y.

Consequences of Non-Linearity

The parameters in the regression model we discuss are the least squares (OLS) estimators of the population regression parameters. These parameters are stochastic variables with a distribution. The assumption of linearity is used to determine the mean of the distribution of these stochastic variables. Only when the assumption of linearity holds will the expected value of the parameter equal the population value of the parameter. When linearity is violated, estimates of the regression parameters can be biased.

When we specify a linear regression model in which the relationship between the dependent variable Y and the independent variables X is not linear, a specification error is made. In that case, the model is not appropriate to describe the dependency between Y and X. Notice that it is always possible to calculate the regression parameters of a model when the assumptions are not fulfilled. However, in that case the estimated parameters of the regression model will be biased.

Diagnostics

To check the assumption of linearity, a statistical test and a few graphical methods can be used.

Statistical Test: The Lack-of-Fit Test

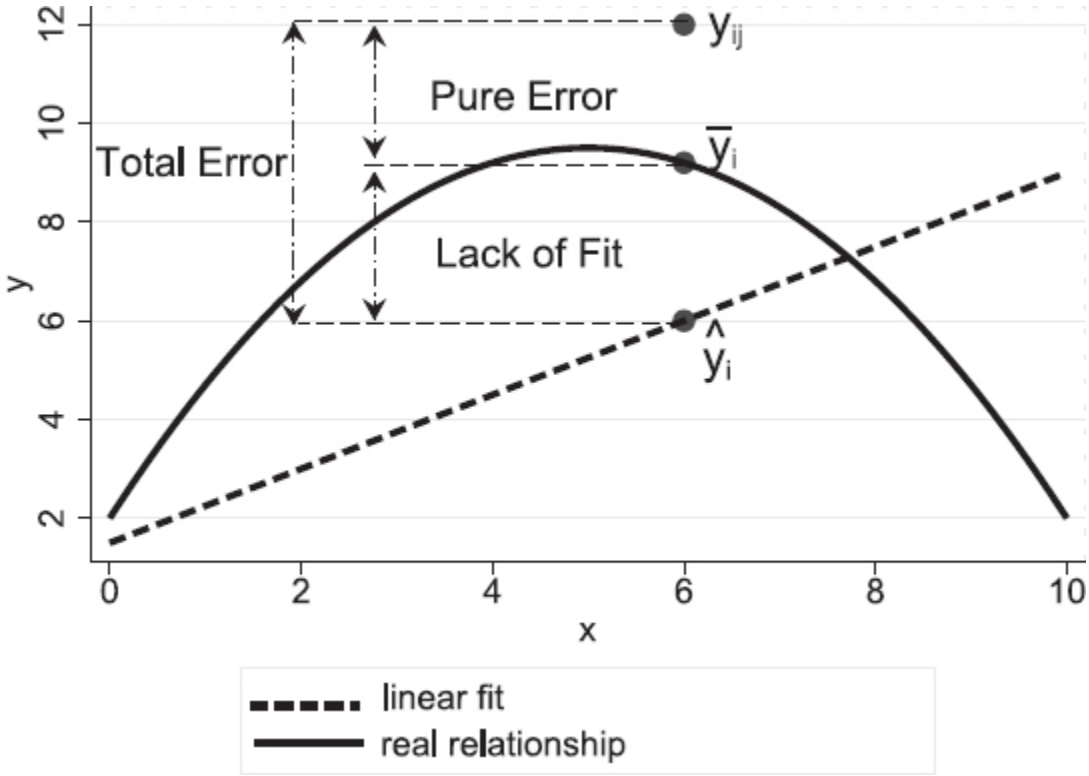
To test the linearity assumption one can partition the error sum of squares (SSE) of the estimated regression model into a ‘pure' error sum of squares (SSE') and lack-of fit sum of squares (SSLF). The ‘pure' error sum of square is the variation of the observed values of Y around their conditional mean (i.e. the mean given a

particular value of *X*). The lack-of-fit sum of squares is the variation of the conditional mean values around the prediction based on the linear model. As such, the SSLF represents deviations from linearity. The basic structure of this partitioning is: SSE = SSE' + SSLF. Or, more fully elaborated:

$$\sum_i \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_i)^2 = \sum_i \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 + \sum_i n_i (\overline{Y}_i - \hat{Y}_i)^2 . \quad (5.2)$$

Figure 5.1 illustrates this idea by means of a simple example. Imagine that the relation between *X* and *Y* is modelled as a linear function (i.e. the dashed line). In reality, however, the conditional means do not lie on a straight line, but follow a quadratic function (the solid line). Now, the distance between observed value Y_i and predicted score $\hat{Y}_i$ (total error) can be divided into the distance between the conditional mean $\overline{Y}_i$ and the observed score $\hat{Y}_i$ (i.e. the pure error) and the distance between the conditional mean $\overline{Y}_i$ and predicted score $\hat{Y}_i$ (lack of fit). The larger the lack-of-fit sum of squares, the stronger the deviation from linearity.

Figure 5.1 Partitioning error sum of square into 'pure' error sum of square and lack-of-fit sum of squares



To test the linearity assumption formally, the amount of pure and lack-of-fit error can be compared by calculating the ratio of SSLF and SSE'. The resulting test statistic follows an *F*-distribution.[2]

$$F = \frac{\mathrm{SSLF}/(c-p)}{\mathrm{SSE}'/(n-c)} \quad (5.3)$$

SSLF has *c*-*p* degrees of freedom, where *c* is the number of actually existing covariate patterns (i.e. the number of combinations of categories of independent variables for which we have observations in the data);

p is the number of regression parameters. SSE' has $n - c$ degrees of freedom (with *n* equal to the sample size). The null hypothesis of this *F*-test states that SSLF is zero, implying that all error is pure error and that the relationship between *X* and *Y* is linear. The alternative hypothesis is that the relationship is not linear.

By way of illustration, we perform a lack-of-fit test for the model presented in Table 5.1. The test gives an *F*-value of 1.2374 (see Table 5.2). The associated *p*-value (0.0873) is greater than $\alpha = 0.05$, meaning that the assumption of linearity is not violated. Note that, because our model includes several variables with a large number of categories (e.g. age and eduyrs), the number of observed covariate patterns (*c*) is very high (1550).

Table 5.2 Lack-of-fit test output

Statistic	Value	df
pN	1648	
SSLF (df)	5366.5905	(1544)
SSE' (df)	275.2833	(98)
$F\ (df_n, df_d)$	1.2374	(1544, 98)
$p > F$	0.0873	

$^{*}p < 0.05$

$^{**}p < 0.01$

$^{***}p < 0.001$

It should be stressed that the *F*-test is an overall test. When the null hypothesis is rejected, there is evidence that the assumption of linearity is not tenable for at least one independent variable. However, we cannot precisely locate the problem. The graphical methods discussed below are useful tools for identifying the variables that are problematic in this regard.

Graphical Method 1: Scatter Plots with Lowess Curve

Graphical methods can be used as an exploratory tool to get a first idea of the relationship between the dependent and the independent variables and also to evaluate the linearity assumption. In a simple scatter plot with the dependent variable and one independent variable, the data points must show a negative or a positive linear relationship between both variables. To get a better visualization of the trend in the scatter plot, one can superimpose a lowess (locally weighted scatter plot smoother) fit line. The lowess method makes no assumption about the form of the relationship between *Y* and *X* (e.g. linear model) and produces a smooth line that follows the trend in the data. The lowess method successively calculates a predicted value for *Y* using a subset of cases (smoothing window) surrounding each value of *X*. If the lowess curve approximately follows a straight line, the linearity assumption is supported.

Figure 5.2 presents scatter plots with lowess curves for each of the independent variables in the regression

model explaining trust in the police. Deviations between the lowess curve and the linear fit are minimal for all variables. As such, this graphical method confirms the lack-of-fit test. A very detailed look reveals that the association between age and trust in the police is slightly curvilinear, with a reversal of the negative age effect around age 70. However, this deviation from linearity is too small to be substantial. Furthermore, we see an outlying observation for education that strongly bends the lowess curve, and could influence the linear fit. We return to this issue when discussing outliers and influential observations later in this chapter.

Graphical Method 2: Residual and Partial Residual Plots

Partial residual plots are a second useful graphical tool to evaluate the assumption of linearity, complementing the information from the scatter plots. In a scatter plot, we get an overview of the marginal relationships between *Y* and *X*. In a multiple regression model, however, our interest lies in the partial relationship between *Y* and *X*, that is, controlling for the other independent variables. Partial residual plots can tackle this problem. Yet before we discuss the partial residual plot we briefly introduce the simple residual plot.

The residual values $e_i = Y_i - \widehat{Y}_i$ can be plotted against each independent variable *X*. This results in a simple residual plot for a particular independent variable. The residuals are represented on the vertical axis, while the values of the independent variable are plotted on the horizontal axis of the graph. In the graph a horizontal line is drawn where the residuals equal zero (the 0-line). This line represents the situation where there is no difference between observed and predicted values of *Y*. When the relationship between *Y* and *X* is properly specified, the points should be scattered randomly around this line, not showing a systematic pattern. This means that the values of *X* are not systematically related with positive or negative residuals; predictive values are not systematically higher or lower than the observed values for particular values of *X*. Once again, a lowess line can be used to visualize the trend. Linearity implies that a'lowess line' approximately follows the 0-line.

Notice that in a multiple regression model the predicted values and, as a consequence, the residuals are determined by several independent variables. As a result, residual plots do not make it possible to link deviations from linearity to a particular independent variable. A partial residual plot can solve this problem. In a partial residual plot an adjusted dependent variable is used:

$$Y_i - \sum_{k,k\neq j} \beta_k X_{ik} = \beta_j X_{ij} + \epsilon_i. \qquad (5.4)$$

In the adjusted dependent variable the linear effects of all independent variables except one (X_j) are subtracted. This means that the dependent variable is corrected for the linear effects of the dependent variables, except X_j. The values of the adjusted dependent variables are called the partial residuals. These values contain two components: the linear effect of the independent variable X_j, and the residual values. For this reason, the partial residual plot is sometimes also called a ‘component-plus-residual plot'. It is the partial residuals that should be linearly related to X_j. A partial residual plot is appropriate to evaluate this relationship. In this plot the *y*-axis is identified by the partial residuals and the *x*-axis by the independent variable X_j. If

linearity holds, the data points in this plot should follow a straight line. Once again, one can plot a lowess curve and a linear fit line to get a clearer visualization. Deviations between the lowess curve and fit line are indicative of deviations from linearity in the partial relationship between *X* and *Y*.

Partial residual plots for the model explaining trust in the police are shown in Figure 5.3. These plots are similar to the scatter plots presented in Figure 5.2, but whereas the plots in Figure 5.2 show bivariate associations, those in Figure 5.3 show partial associations. None of the plots give evidence for violations of the linearity assumption, as there are no substantial differences between the lowess curve (the dashed line) and the linear fit line (the solid line). In this case, the conclusions from the scatter plots and partial residual plots are identical, but this need not always be the case.

Figure 5.2 Combined scatter plots with linear fit line and lowess curve

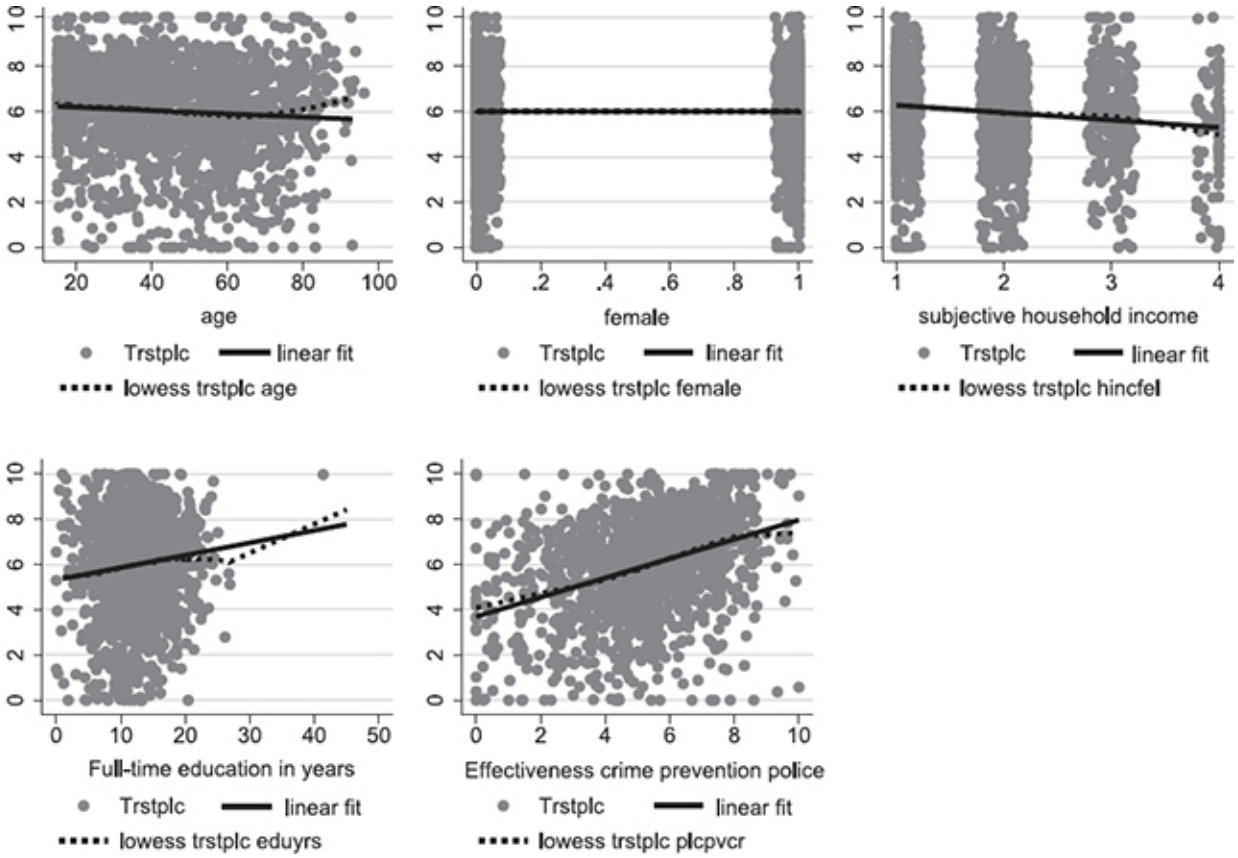
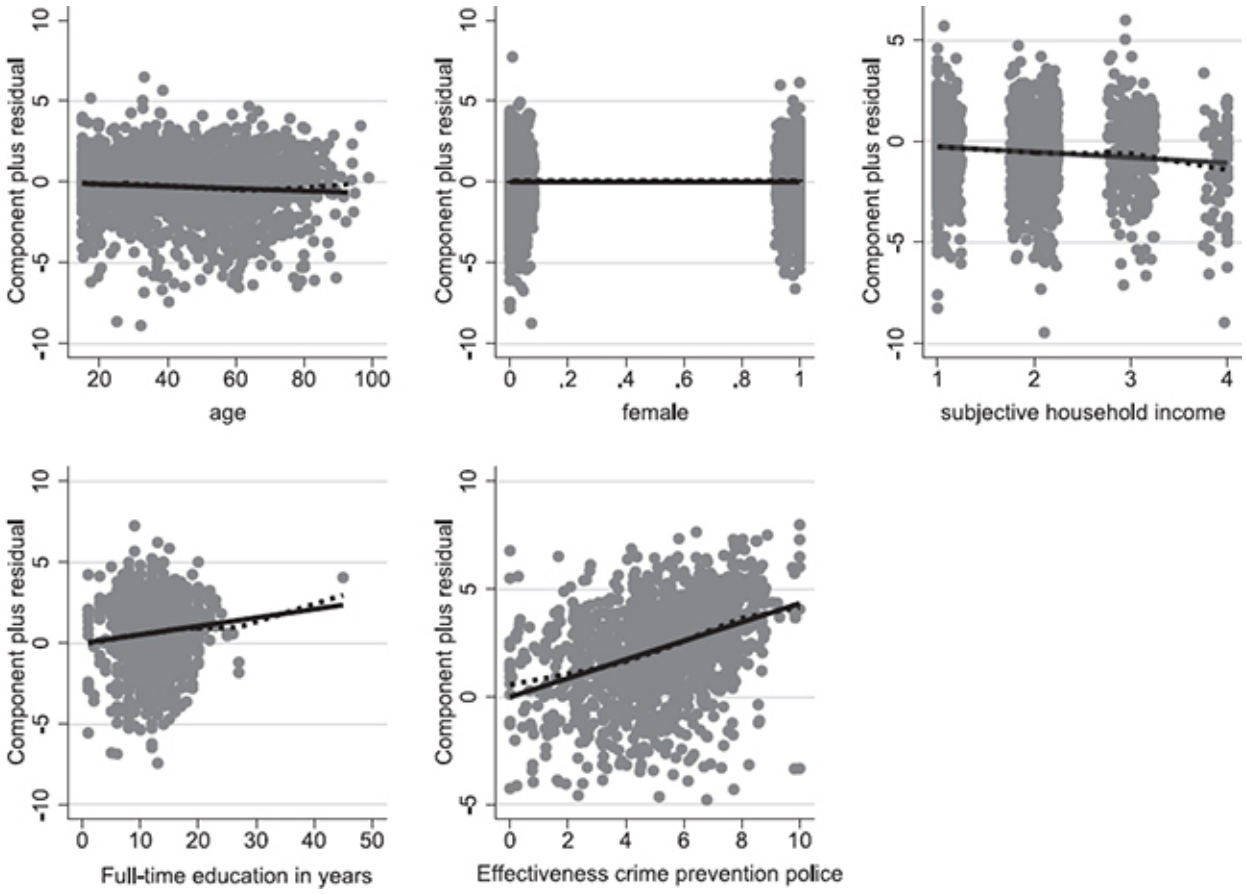


Figure 5.3 Partial residual (component-plus-residual) plots x



Remedies

As we mentioned before, violations of the assumptions of linearity can have severe consequences and may lead to biased estimates. When non-linear relationships between *Y* and *X* are detected, it is recommended to use procedures that account for the non-linearity. The literature on modelling non-linearity is extensive. In this presentation, we restrict ourselves to a brief explanation of two popular approaches: polynomial regression and piecewise regression. A more detailed account of polynomial regression can be found in Chapter 6 of this volume, while piecewise regression is discussed extensively in Chapter 14.

Both procedures can be considered as an adjustment or manipulation of the independent variable(s). The independent variable(s) are transformed in such a way that the relationship between the transformed independent variable and the dependent variable follows the structure in the data.

Polynomial Regression

In a polynomial regression model, second- or higher-order terms are entered into the model. These terms are powers of the *X* variable and they serve as additional predictors. When the relationship between *X* and *Y* is not linear, several higher-order terms can solve the non-linearity. When, for example, the relationship between *X* and *Y* is curvilinear with one maximum, the appropriate model is a quadratic regression model

(i.e. a second-order polynomial in X_1): $\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2$. In this model, there is only one substantive independent variable: X_1. A positive β_2 in the quadratic model indicates a model that is U-shaped (concave upwards); a negative indicates a curve that is inverted U-shaped (concave downward). One can elaborate the polynomial model with several higher-order terms. In a cubic model, for example, we introduce a third-order polynomial in X_1 and we get $\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i1}^3$. The number of higher-order terms needed in the model depends on the strength of the non-linear relationship between X and Y. Notice that polynomial regression models are still linear models, because they are linear in their parameters.

Piecewise Regression

Sometimes the non-linear curve is not smooth, but characterized by one or several breaking points. In that case, the relationship between dependent and independent variables is different for different segments of the range of the independent variable. The segments are delimited by breaking points where the effect of the independent variable substantially changes.

Take, for example, a simple regression model with one breaking point and where the relationship between the independent and the dependent is linear before and after the breaking point. In a piecewise regression model for this situation, the range of X is divided into two segments, $X < b$ and $X < b$, where b is the value of X at the breaking point in the regression line. To solve the problem of non-linearity we need a separate regression equation for each segment. Therefore, we define and use a new independent variable: $X_B = 0$ if $X < b$, and $X_B = X < b$ if $X \geq b$. This new variable is entered into the following regression equation:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_b X_{iB} + \epsilon_i. \qquad (5.5)$$

The regression coefficient of the new independent variable X_B is now the change in the slope of the regression line used before the breaking point.

Assumption 2: Homoscedasticity or Constant Variance Assumption

What Is It?

The assumption about homoscedasticity is related to the dispersion of error terms or residuals of the model. The assumption is that for each covariate pattern of X, the variance of the residuals is constant, $\mathrm{Var}(\epsilon_i | x_{i1}, \ldots, x_{ik}) = \sigma^2$. This means that, when the residuals are uncorrelated (see the independence assumption), the variance–covariance matrix of the residuals can be written as

$$\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}. \qquad (5.6)$$

The conditional variances of the residuals represent the variability of the residuals around the predicted value based on a specific combination of values of independent variables X. So homoscedasticity means that all

the conditional residual variances are equal: residual variances are constant regardless of the values of the independent variables *X*.

Heteroscedasticity conversely refers to the situation of non-constant variance of the residuals. The variance of the residuals changes as the value of *X* changes. Heteroscedasticity or the dependency of the variances of the residuals and the values of *X* can occur in different ways. It is possible, for example, that the variance of residual values increases when the predicted value of *Y* increases. This means that the predictions based on the model are better for low predicted values of the dependent variable than for high predicted values. The reverse situation is also possible: a decrease in the variances when the predicted values of *Y* increase. More complex patterns of heteroscedasticity can also occur.

Consequences of Heteroscedasticity

The assumption of constant error variance is used to determine the variance of the distribution of the parameters in the standard OLS estimation procedure of a regression model. Faulty inferences can be made when this assumption does not hold. Remember that the OLS procedure is used to estimate the parameters **b** (vector of the estimated regression parameters) and **V**(**b**) (the variance–covariance matrix of the regression parameters). In matrix notation, this can be written as

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{5.7}$$

$$\mathbf{V}(\mathbf{b}) = \mathbf{V}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \tag{5.8}$$

Now, when **V**(**y**) is assumed to equal $\sigma^2\mathbf{I}$ (thereby implying constant error variance), the expression for **V**(**b**) simplifies greatly to:

$$\mathbf{V}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{5.9}$$

Yet, when the error variance is not constant in the data, the calculation of the estimated standard errors of the regression parameters is no longer valid. In this situation, the least squares estimators are still unbiased, but they are inefficient. This means that when non-constant error variance occurs, there exist other estimation procedures (weighted least squares; see the subsection below on remedies) that generally produce estimators with smaller standard errors. It is clear that this has an impact on the significance tests and confidence intervals.

Diagnostics

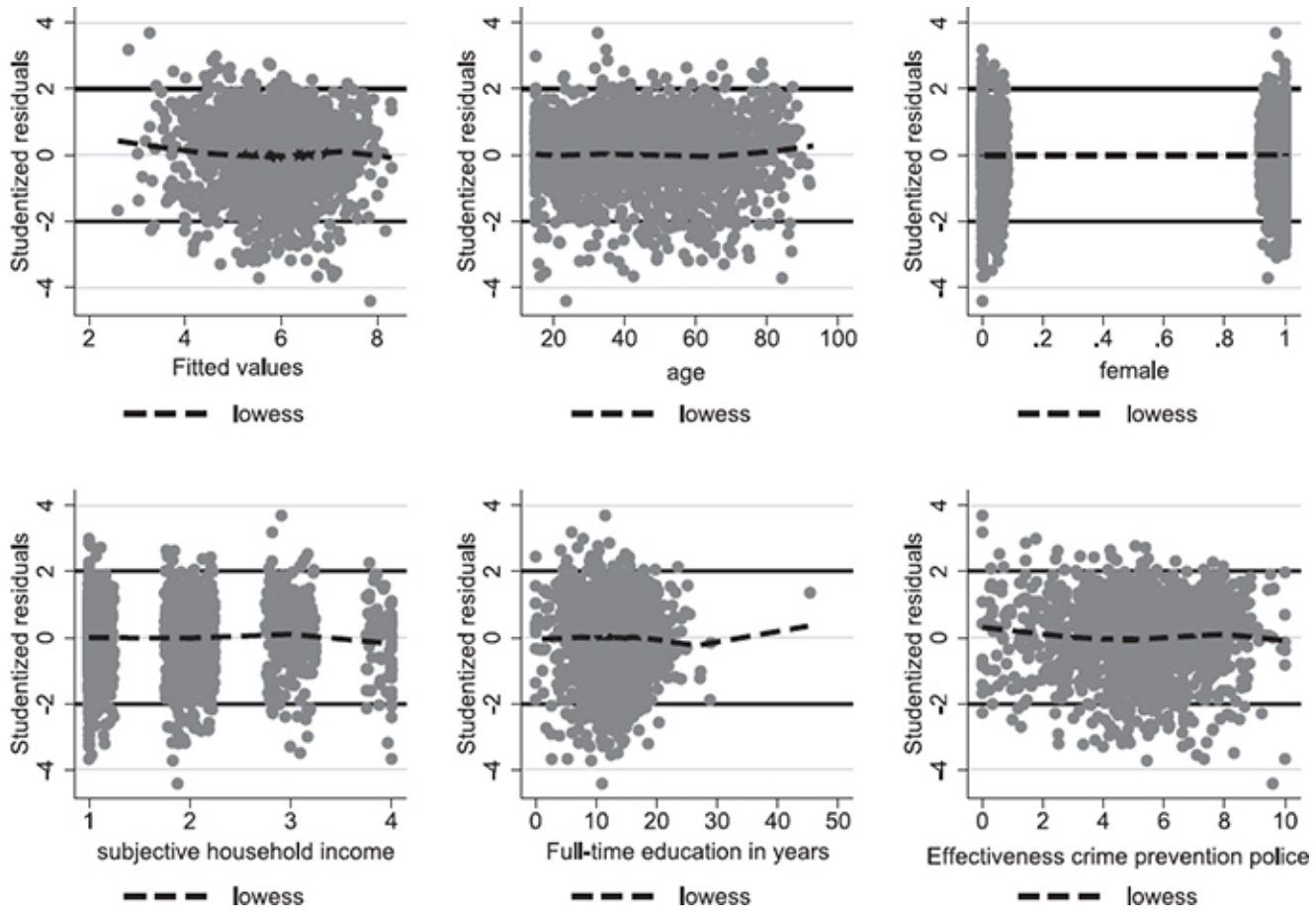
Graphical Method: Studentized Residual Plots

Once again a graphical tool is useful to evaluate the homoscedasticity assumption. The most appropriate graph is a plot of studentized residuals against the predicted values of *Y* or against certain predictors. Studentized residuals are residuals divided by an estimate of their variance (this variance is unknown, and therefore it is impossible to standardize). Concretely, residuals are studentized by dividing them by

$s_i^2\sqrt{1-h_i}$ (where s^2_i is the estimate of σ^2 obtained after deleting the ith observation and h_i is the leverage of i – see the section on influential observations for more information on leverage). Plotting studentized rather than raw residuals makes it easier to observe patterns of changing spread (Fox, 2008, p. 272). When the studentized residuals are randomly spread around the mean of zero and contained within a horizontal band of ±2 standard deviations of the mean, the homoscedasticity assumption can be considered valid. A pattern of changing dispersion in the studentized residuals (e.g. increase or decrease of the spread with the level of the predicted values of *Y* or one of the *X*s) is indicative of heteroscedasticity.

Figure 5.4 displays studentized residuals for the regression model explaining trust in the police. Studentized residuals are plotted against fitted values (upper left-hand corner) as well as against the various predictor variables in the model. The plot clearly reveals some anomalies. High studentized residuals (> 2) occur more often for low fitted values, while low studentized residuals are mostly present when fitted values are high. This pattern suggests that heteroscedasticity is present. A similar pattern can be seen for the plot for predictor plcpvcr (effectiveness of crime prevention).

Figure 5.4 Residual versus fitted plot (upper left) and residual versus predictor plots



Statistical Test: White's Test

White's test is a formal procedure for detecting heteroscedasticity (White, 1980). Although this test is a more general test for model misspecification, it is also appropriate for testing homoscedasticity. The test does

not make any assumption about the pattern of non-constant error variance. The null hypothesis is that the residuals are homoscedastic:

$$H_0 : \sigma^2_{\epsilon_i} = \sigma^2_{\epsilon}. \qquad (5.10)$$

The rejection of this null hypothesis is evidence of heteroscedasticity. In White's test, the squares of the residuals of a substantive model with *p*-1 predictors (step 1) are regressed on all predictors in the model plus all cross products among the predictors (step 2). The test statistic in White's test equals nR^2 of the last model (step 2). Under the null hypothesis of homoscedasticity, the test statistic is distributed as chi-squared with degrees of freedom the number of predictors of the model in step 2.

In our example, White's test renders a χ^2-value of 116.13 (for 19 degrees of freedom). This value is strongly significant ($p < 0.0001$) and confirms that a significant amount of heteroscedasticity is present in the data.

Remedies

Several strategies can be used to tackle the problem of heteroscedasticity. A first remedy is a transformation of the dependent variable. When, for example, the spread of the residuals is an increasing linear function of the predicted values one can use the square root of the dependent variable. Alternatively, a log or inverse transformation can be used.

Another frequently used strategy to deal with heteroscedastic data is to perform a weighted least squares (WLS) estimation procedure instead of ordinary least squares. Remember that in an OLS procedure the values of the parameters of the model are estimated by minimizing the value of the sum of the squared residuals: min($\sum e_i^2$). In the OLS procedure, all units have the same weight, $w_i = 1$. In a WLS estimation procedure, each unit is given a different weight w_i and the sum of the weighted squared residuals is minimized: min($\sum w_i e_i^2$).

The generalized least squares procedure is the starting point for tackling heteroscedasticity. In this generalization of OLS, the inverse of the diagonal matrix Var(ε_i) = **V** is used. We get

$$\mathbf{b}_w = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \qquad (5.11)$$

We obtain $\mathbf{V}^{-1}$ by inverting the diagonal elements of **V**. This results in D_{wi}, a diagonal matrix with $w_i = 1/\sigma^2_i$ on the diagonal and $b_w = (X'D_{wi}X)^{-1} X'D_{wi}y$. So, to treat the heteroscedasticity problem, the inverse of the conditional variance of the residuals is used as weight for each unit: $w_i = 1/\sigma^2_i$. In this way, units with larger residual variance are given a smaller weight than units with less variance. One can also consider the conditional variance of the residuals as an indicator of the precision of the estimate in that condition. Units with less precision (large variance) are given smaller weights than units with more precision.

It can be proven that when **X** and **y** are multiplied by $\sqrt{w_i}$, an OLS regression with these transformed

variables results in **b***w*, and that this process is equivalent to minimizing the weighted sum of squares: min($\sum e_i^2/\sigma_i^2$).

To illustrate the functioning of this weighting procedure, take a regression model with one independent variable *X* and heteroscedastic residuals (Panik, 2009, p. 157). We assume that the other assumptions of the regression model are correct. The basic expression for the model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \quad \text{Var}(\epsilon_i) = \sigma_i^2,\ i = 1, \ldots n. \tag{5.12}$$

After transforming this model by multiplying both sides by $1/\sigma_i$, we get the expression

$$\frac{Y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{X_{i1}}{\sigma_i} + \frac{\epsilon_i}{\sigma_i}, \quad \sigma_i \neq 0,$$

or

$$Y_i^* = \beta_0 W_i^* + \beta_1 X_{i1}^* + \epsilon_i^*, \tag{5.13}$$

with

$$Y_i^* = \frac{Y_i}{\sigma_i}, \quad W_i^* = \frac{1}{\sigma_i}, \quad X_{i1}^* = \frac{X_{i1}}{\sigma_i}, \quad \epsilon_i^* = \frac{\epsilon_i}{\sigma_i}.$$

For the residuals of this model, we have

$$E(\epsilon_i^*) = \frac{E(\epsilon_i)}{\sigma_i} = 0,$$
$$\text{Var}(\epsilon_i^*) = \frac{\text{Var}(\epsilon_i)}{\sigma_i^2} = 1. \tag{5.14}$$

This means that in the transformed model, the conditional variance of the residuals is constant. Notice that we obtain this result by multiplying the dependent and independent variable by the square root of the weights and these weights are the reciprocals of the residual variances. When it is possible to assume that the conditional variance of the residuals is a function of an independent variable *X*, we get other weights and another transformation of the variables in the model. Suppose, for example, $\sigma^2_i = \sigma^2 X_i^2$; then $w_i = 1/\sigma^2 X_i^2$ and we transform the model by multiplying both sides by $1/X_i$. Then the residual equals

$$\epsilon_i^* = \frac{\sigma \epsilon_i}{\sigma_i}, \quad \text{and} \quad \text{Var}(\epsilon_i^*) = \frac{\sigma\ \text{Var}(\epsilon_i)}{\sigma_i^2} = \sigma. \tag{5.15}$$

Once again we obtain constant error variance.

In practice, the variance of the residuals is not known and must be estimated from the data. It can be shown that squared observed residuals (or the absolute values of the residuals) are unbiased estimates of the population variance. However, the reciprocal of the squared observed residuals cannot directly be used as weight in the WLS procedure. After all, units with the same covariate pattern (same values on the independent variable) do not always have the same value for the dependent variable *Y*. As a consequence,

they have a different (squared or absolute) residual value and this results in a different weight. However, we need the same weight for units with the same covariate pattern. This can be realized using the following procedure. We first specify a regression model with a substantive dependent (**y**) and independent variables (**X**). This regression analysis produces a residual value for each unit. Second, to get the same weight for each unit with the same covariate pattern, the squared residuals (or the absolute values of the residuals) are used as the dependent variable in a model with the relevant substantive variables which sufficiently explain the residuals. To avoid negative weights it is better to use the log of the squared residuals (log e^2_i). This regression model results in the log of predicted squared errors ($\log \widehat{e_i^2}$) which are equal for all the units with the same covariate pattern. To recover the estimated squared residuals ($\widehat{e_i^2}$), the inverse log is taken. In the last step the WLS estimators are produced by an OLS regression analysis with the transformed variables. The weights used to transform the variables are $w_i = 1/\widehat{e_i^2}$. The stepwise summary of the procedure following DeMaris (2004, p. 206) is:

1. Regress *Y* on *X* and save e_i.
2. Transform e_i into log e^2_i.
3. Regress log $e_{2\ i}$ on *X* and save the predicted values log$\hat{e}^2_i$.
4. Take the inverse of log $\hat{e}^2_i$ to get $\hat{e}^2_i$.
5. Regress *Y* on *X* using $w_i = 1/\hat{e}^2_i$.

It is important to notice that the reported R^2 on the output of WLS regression analysis is calculated for the transformed data and not for the original data. Because of this, it is not valid to compare the R^2 of both analyses. To obtain a comparable value of R^2 we must use the original data and the WLS estimates and calculate the predicted values and the WLS residuals. Then the sum of squared WLS residuals is used to calculate $R2_{WLS}$. Usually this value is not reported by the software.

A final small comment is related to the use of sampling weights. Sampling weights are used to correct for different probabilities of selection into the sample and to make sample and population distributions comparable. Sampling weights are different from the weights used in the WLS procedure. These weights are a function of the error variances and used to produce correct estimates of the standard error. It can be shown that using sampling weights produces heteroscedasticity even when the unweighted data are homoscedastic.

Table 5.3 compares the OLS estimates from Table 5.1 with results obtained through WLS estimation with $w_i = 1/\widehat{e_i^2}$ as weights. As expected, the WLS procedure generally leads to slightly decreased standard errors, while parameter estimates do not differ substantially. Yet the differences between OLS and WLS are not dramatic. This is not surprising, since the studentized residual plots revealed relatively small amounts of heteroscedasticity.

Table 5.3 OLS and WLS regression output

	OLS		WLS	
Variable	*b*	*SE*	*b*	*SE*
Constant	3.819	(0.287)	3.398	(0.246)
Age	–0.007	(0.003)	–0.005	(0.002)
Female	0.003	(0.092)	–0.051	(0.086)
Hincfel	–0.268	(0.056)	–0.249	(0.054)
Eduyrs	0.052	(0.013)	0.067	(0.010)
Plcpvcr	0.434	(0.025)	0.460	(0.026)
Adjusted R^2	0.177		0.190	

Note: Analytic weights = 1/exp (*g*).

Assumption 3: Independence of Residuals

What Is It?

As a third assumption, the regression model requires independence of the error terms. The residuals should be patternless, meaning that the residual value for one observation cannot depend on the residual for other data points. Formally, this means that the residual values should not be correlated:

$$E(\epsilon_i, \epsilon_j) = 0 \quad \text{for all } i \neq j. \tag{5.16}$$

The independence assumption can also be looked at from the perspective of the dependent variable. It implies that, conditional on the independent variables, the values Y_i(with $i = 1, \ldots, n$) should be independent draws from the population distribution of Y.

Violations of independence occur when important explanatory factors are neglected in the model. Think of a dependent variable, such as personal income, for which important gender differences exist (males earning more than females). If gender is omitted from the regression model explaining income, we tend to underestimate male incomes systematically, and overestimate female incomes. Males, on average, will have positive residuals and females negative ones. As result, the residuals are patterned: for a gender-matched couple of respondents, a positive correlation between the error terms is expected; for a male and a female respondent, the correlation between residuals will be negative. Note, however, that this pattern will be hard to detect, since in this example gender is not a variable in the model.

In practice, non-independence mainly occurs in two situations. The first situation is time series data, where some dependent variable is measured several times on different occasions. Since social phenomena tend to change only gradually over time, the observation at time point t will probably depend to some extent on how the situation was at time point t −1. In this case, the data is said to be autocorrelated: the correlation for residual values of consecutive observations is not zero. Clustering of observations is a second common violation of non-independence. In many instances, social scientists encounter hierarchically structured data:

observations are nested in higher-level units. This clustering is often a result of the fact that social reality is layered. Persons are not atomized individuals, but instead grow up in families, work together in organizations and live together in neighborhoods. But clustering can also arise from the research design (e.g. when groups of respondents are interviewed by the same interviewer). If cluster membership is related to the dependent variable but not taken into account into the model, observations belonging to the same cluster will have more similar residuals as observations belonging to other clusters. Again, the assumption of independent error terms is violated.

Consequences

Violations of the independence assumption affect the standard errors of the regression parameters rather than the parameter estimates themselves. As a result, non-independence can affect statistical inference, but does not invalidate regression analysis as a descriptive tool.

The consequences of non-independence for statistical inference can be illustrated in an intuitive way using the following example. Imagine a study in which 100 monozygotic twins are investigated. Obviously, the independence assumption does not hold here. Twins share their genetic material as well as important life experiences. As a result, twin siblings are more similar than two respondents who are not siblings, and will consequently have similar residuals. Yet because of this similarity, a pair of twins contains less information than two independent individuals do. Including someone's twin brother or sister in the study adds relatively little new empirical evidence, because in large part the same information is replicated. Including a completely unrelated individual instead contributes a larger amount new information. Thus, although our hypothetical twin study contains 200 respondents, the data is less informative than a study containing 200 unrelated persons. Nevertheless, the regression model assuming independence proceeds as if 200 independent observations are present, overestimating the amount of available information and consequently also the reliability of the estimates. As a result, the estimated standard errors tend to be smaller than they are in reality. This can result in pseudo-significant effects.

The same point can be made in a more formal manner as well. The independence assumption is used in the derivation of the variance of the regression coefficients. To simplify matters, take the example of simple linear regression with one independent variable. In this case, the point estimator of the regression coefficient can be written as (Neter et al., 1996, p. 46):

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\,Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad (5.17)$$

Consequently, the variance of this point estimator equals

$$\mathrm{Var}(b_1) = \mathrm{Var}\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})\,Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]. \qquad (5.18)$$

Regression theory considers the values of the predictor *X* as known constants. As a result, the observations

Y_i(keeping X constant) are the only random variables in this expression, and Var(b_1) can be rewritten as

$$\mathrm{Var}(b_1) = \frac{\mathrm{Var}\sum_{i=1}^{n}(X_i - \overline{X})\,Y_i}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}. \qquad (5.19)$$

The formula above expresses the variance of estimator b_1 as the variance of a (weighted) sum of random variables Y_i. A fundamental law of variances and covariances states that the variance of a sum of two random variables equals the sum of the variances of the random variables plus two times its covariance:

$$\mathrm{Var}(A + B) = \mathrm{Var}(A) + \mathrm{Var}(B) + 2 \times \mathrm{Cov}(A, B). \qquad (5.20)$$

The fact that covariances between all the Y_i need to be taken into account makes the further elaboration of Var(b_1) a very complex operation. Assuming that the covariances of the random variable Y_i(conditional on X) equal 0 simplifies calculation considerably. Under this condition, the variance of a sum of the random variables can be written as the sum of its variances:

$$\frac{\mathrm{Var}\sum_{i=1}^{n}(X_i - \overline{X})\,Y_i}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2} = \frac{\sum_{i=1}^{n}\mathrm{Var}\left[(X_i - \overline{X})\,Y_i\right]}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}. \qquad (5.21)$$

Now that the covariances have been removed as obstacles, and assuming that all the Y_i have the same variance σ^2 (i.e. homoscedasticity – see assumption 2), the variance of the regression coefficient can be easily obtained:

$$\mathrm{Var}(b_1) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}\mathrm{Var}(Y_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}. \qquad (5.22)$$

If in reality the covariances of the Y_i(conditional on X) are not zero while we assume them to be, the variance of the regression coefficient will be incorrectly calculated. The harmful effects for statistical inference can be serious. Scariano and Davenport (1987), for example, show by means of a simulation study that even mild clustering can produce pseudo-significant effects, since probabilities of falsely rejecting a null hypothesis of an F-test for equal means are inflated.

Diagnostics and Remedies

Earlier, we mentioned that non-independence most often occurs in the case of time series or nested data. For both situations, diagnostic tools as well as specific models taking the non-independence into account have been developed. Because of their importance, these models are discussed in greater detail elsewhere in this volume. Chapter 7 explains how clustered data can be properly analyzed using multilevel models. Models for time series data are discussed in Chapter 17.

Assumption 4: Normality

What Is It?

The standard regression model presupposes that the distribution of the residuals or error terms e_i has a particular form; the error terms are assumed to follow a normal distribution. Concretely, the density function of residuals should have the famous bell shape of a Gaussian curve described by the following function:

$$f(\epsilon_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \qquad (5.23)$$

where μ is the mean of the residuals (which equals 0 by definition in regression analysis) and σ^2 is the variance of the residuals. This implies, among others, that the distribution of the residuals should have a single peak (i.e. be unimodal), be symmetric instead of skewed (skewness = 0), and that the tails of the distribution should be neither too light nor too heavy (kurtosis = 0).

The normality assumption has repercussions for the distribution of the dependent variable as well. If we consider values of the predictor variables that are considered as fixed (as regression theory does), the observations Y_i and residuals e_i are communicating vessels. The assumption of the normal residuals implies normality for observations of the dependent variable Y_i conditional on the independent variables X (i.e. for specific values of the X's). This is, however, not necessarily the same as saying that the marginal distribution of dependent variable Y should be normal.

One can easily imagine a dependent variable Y following a normal distribution for males and females separately. If these gender-specific normal distributions have a different mean, however, the distribution of Y in the population of men and women together can be bimodal and thus not normal. Likewise, a normally distributed dependent variable does not guarantee normally distributed residuals. It is the dependent variable Y, controlling for the independent variables, that needs to fulfill the requirement of normality (Lumley et al., 2002). Nevertheless, the normality assumption bears some implications for the distribution of the dependent variable Y as well. The residuals can only fully match the normal density function if Y is continuous (rather than categorical) and not truncated (i.e. not having a lower or upper limit).

Consequences of Non-Normality

The normality assumption places rather strict restrictions on the residuals, and will very often be violated in social science research. Fortunately, the consequences of non-normality range from quite mild to even non-existent. The normality assumption is used to determine the functional form of the sampling distribution of the regression estimates. To illustrate this, consider the case of a linear regression model for Y_i with only one predictor variable X. We have seen before that the point estimator of the regression slope (b_1) can be written as

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\,Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}. \qquad (5.24)$$

Regression theory considers the values of the predictor X as fixed. As a result, the observations Y_i(keeping X constant) are the only random variables in the expression for regression estimate b_1; b_1 is a linear combination of observationsY_i. If we assume that these observations Y_i, conditional on X, are normally distributed – in addition to being independent (see assumption 3) – the sampling distribution of b_1 is also normal (Neter et al., 1996, p. 1320). The mean of this normal distribution is the population regression slope β_1 and the variance equals σ^2_{b1}:

$$b_1 \sim N(\beta_1, \sigma^2_{b_1}). \qquad (5.25)$$

Or the standardized regression slope is a standard normal variable (i.e. with mean 0 and standard deviation 1):

$$\frac{b_1 - \beta_1}{\sigma_{b_1}} \sim N(0, 1). \qquad (5.26)$$

Because the population standard deviation σ_{b1} is unknown, however, it needs to be estimated by S_{b1} (see Chapters 2 and 3 in this volume). If we divide by the estimate S_{b1} instead of by the population value σ_{b1}, we get a so-called studentized statistic, following a t-distribution with $n - p$ degrees of freedom:

$$\frac{b_1 - \beta_1}{S_{b_1}} \sim t_{(n-p)}, \qquad (5.27)$$

where p refers to the number of regression parameters (i.e. the number of predictors plus one intercept).

Thus, the normality assumption allows us to derive how the regression estimates are distributed. This information is used to perform significance tests for regression coefficients and to constructed confidence intervals. Consequently, statistical inference regarding the regression coefficient could go off the rails if normality of the residuals does not hold. In that case, the regression parameters (divided by their estimated standard error) are not guaranteed to follow a t-distribution. The reported t-values and their corresponding p-values, as well as the constructed confidence intervals, can be biased.

In practice, however, the consequences of non-normality are often limited. First, the normality assumption is not necessary for point estimates of the regression parameters to be unbiased, and thus has no consequences for regression as a descriptive tool. Second, statistical tests for regression parameters are quite robust against deviations from normality. If sample sizes are sufficiently large, the regression coefficients will approach a t-distribution even if the observations Y_i are not normally distributed. This can be explained by the central limit theorem, which asserts that the sum of a large number of random variables (here, the observations Y_i) will be approximately normally distributed irrespective of the distribution of these random variables. The data sets used in applied research are usually sufficiently large to provide a reasonable approximation (Lumley et al., 2002).

Therefore, some authors argue that practical researchers can neglect the normality assumption (see Gelman and Hill, 2007, p. 46). To a certain extent, we agree with this relaxed position on the normality assumption. This does not mean that normality is completely irrelevant, though. In small data sets with strong violations,

the normality assumption can still be an issue. Furthermore, non-normal data often contain influential observations that can distort the regression analysis (see assumption 5). For that reason, we offer a brief discussion on diagnostics and remedies for non-normality below.

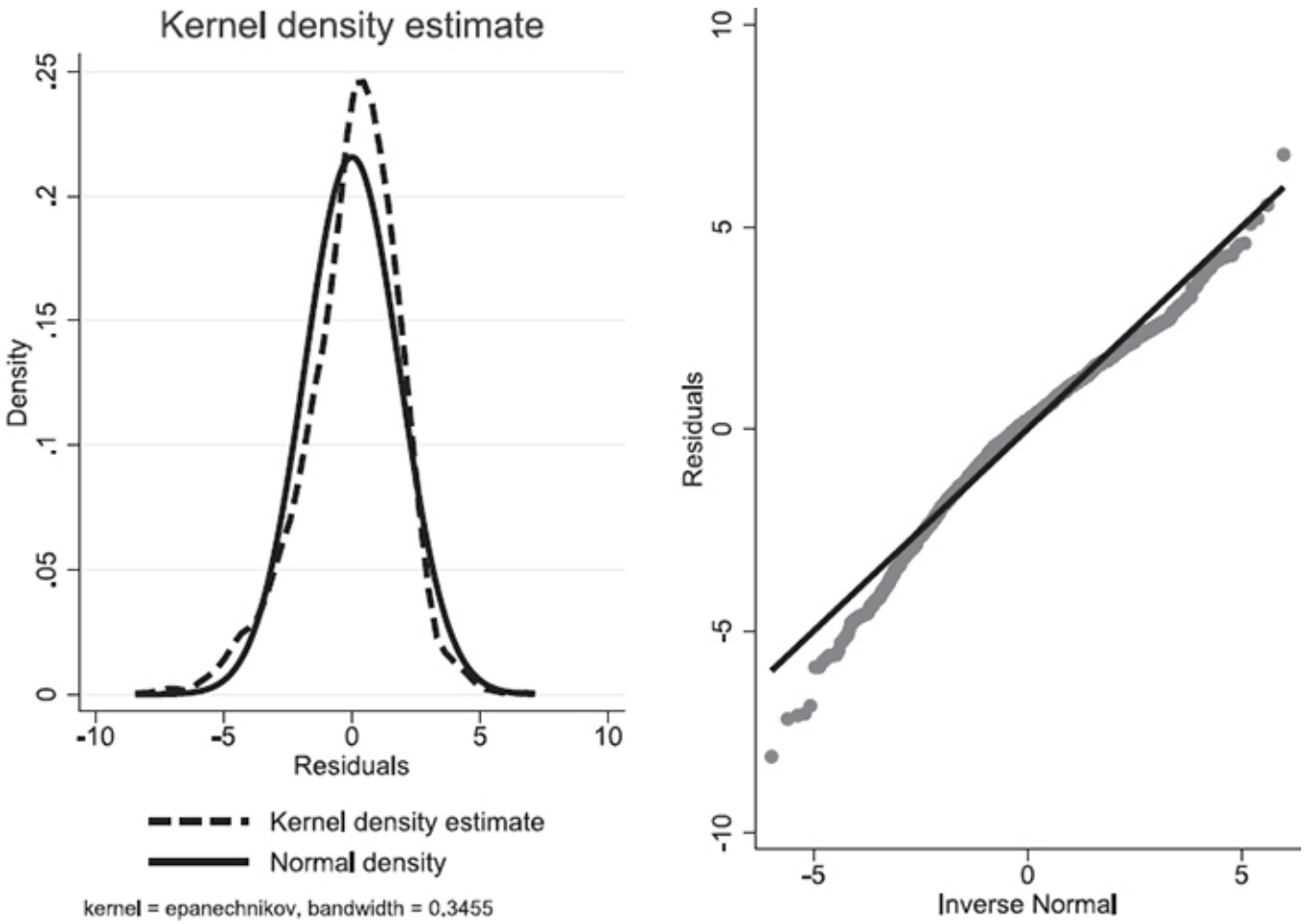
Diagnostics

Graphical Method: The Quantile–Quantile Plot

Normality of the residuals can be assessed by drawing a so-called normal quantile–quantile (QQ) plot or normal probability plot. The QQ plot is a standard graphical tool used to compare observed values to a theoretical distribution. In the normal QQ plot, the quantiles of the observed residuals and the quantiles of the standard normal distribution are plotted against each other. To obtain a QQ plot, the residuals obtained from a regression model are ranked from low to high. These values are called the observed quantiles of the residuals, and are plotted on the vertical axis of the QQ plot. Then, every residual is given a rank order i, going from 1 to n (given that there are n observations). These rank orders are then divided by n. The resulting proportional rank (i/n) of a residual indicates the proportion of the observations that have a smaller or equal residual value. Subsequently, the z-score that would cut off exactly the same proportion in a cumulative standard normal distribution is identified. These z-values are the quantiles of the standard normal distribution, and are plotted against the quantiles of the observed residuals (on the horizontal axis of the QQ plot). If the dots in the resulting scatter plot form a straight line, the residuals are normally distributed. Deviations from a straight line are indicative of non-normality. A drawback of the QQ plots is that the decision whether the plot displays a straight line is always arbitrary to a certain extent.

Figure 5.5 shows a normal QQ plot (right) for the regression model explaining trust in the police, as well as a kernel density plot of the residuals (left). Both plots indicate that the residuals follow a normal distribution reasonably well. The kernel density curve maps rather well onto the Gaussian curve, and the dots on the QQ plot more or less follow a straight line. The most important deviation can be found in the left tail of the distribution. The dots below the line on the left-hand side of the QQ plot mean that the negative residuals are somewhat more extreme than expected. In other words, the left tail is slightly heavier than expected. Yet these are only minor deviations and the main conclusion from the plots is that the normality assumption is sufficiently well approximated.

Figure 5.5 Kernel density and QQ plots



Statistical Tests for Normality: Shapiro–Wilk and Kolmogorov–Smirnov

Alternatively, formal statistical tests for normality are available as well. The reader should be aware, however, that the practical use of these tests is often limited. When sample sizes are small (the only condition under which non-normality substantially affects the conclusions), the tests lack power to detect deviations from normality. When sample sizes are large (thus when the non-normality hardly has perceivable consequences), the tests are sometimes overly sensitive, and even detect insubstantial violations of normality. Nevertheless, we briefly discuss two tests that are readily available in most statistical software packages.

Shapiro and Wilk (1965) developed a statistic, *W*, that quantifies the straightness of the line formed by the dots in a QQ plot. *W* ranges from 0 to 1, where higher values are indicative of a closer correspondence to a normal distribution. If *W* falls below a certain critical value, the null hypothesis of normality is rejected.

The Kolmogorov–Smirnov test evaluates the discrepancy between the cumulative density function of the observed residuals and the cumulative density function of a normal distribution with the same mean and variance. If the residuals are normally distributed, both functions are expected to be highly similar. The Kolmogorov–Smirnov test is based on test statistic *D*, the maximum distance between both functions. A *p*-value is given, representing the probability that an even larger distance is found in the sample, assuming

both density functions are equal in the population. If the *p*-value is smaller than 0.05, then the null hypothesis that the cumulative density functions are equal at all points is rejected, and we conclude that normality does not hold.

Of these two tests, the Shapiro–Wilk test has the larger statistical power (Razali and Wah, 2011). Therefore, it is preferable to use Shapiro–Wilk when sample sizes are small ($n < 100$). For larger sample sizes, Kolmogorov–Smirnov is to be preferred.

Table 5.4 shows both tests for the example regression model. Despite the minor deviations from normality we saw in the kernel density and QQ plot, both tests have a *p*-value smaller than 0.05 and thus indicate that the assumption of normality is significantly violated. This illustrates that these normality tests are very sensitive to deviations from normality when sample sizes are large.

Table 5.4 Results for the statistical tests for normality

Test	Test statistic and p-value
Shapiro–Wilk	0.977 (0.000)
Kolmogorov–Smirnov	0.068 (0.000)

Remedies

Two main strategies exist if non-normality is deemed to be a problem. First, one can adapt the regression model to the data, and model the non-normality. The generalized linear model is a generalization of the basic regression model that makes it possible to relax the assumption of normality and assume other error distributions instead. This makes it possible to model variables following binomial (logistic regression – cf. Chapter 8 in this volume), multinomial (cf. Chapter 9), count, Poisson and numerous other distributions. The alternative strategy consists of adapting the data to the model by transforming the dependent variable *Y* so that its distribution becomes (approximately) normal. Commonly used transformations include the family of Box–Cox transformations or the logarithmic transformation. Because the ability of these transformations to deal with non-normality is of limited use for applied researchers, we do not discuss them in detail here and refer to Carroll and Ruppert (1988) for more details instead.

Assumption 5: Absence of Influential Observations

The Nature of the Problem and Its Consequences

In contrast to the other assumptions, which refer to relations between variables or error distributions in the complete data set, this assumption deals with the position of specific observations. Data sets sometimes contain a small number of observations that are separated from the rest of the data, in the sense that they have values that deviate strongly from the other observations. Potentially, such ‘extreme cases’ can become

influential observations that affect the results of the regression analysis. An influential observation can be defined as a case 'that alters the value of a regression coefficient whenever it is deleted from an analysis' (Allen, 1997, p. 177). Obviously, the presence of influential observations is not desirable, as this means that the regression results are distorted by a couple of extreme observations.

To understand under what conditions extreme cases can become influential observations, two types of extreme observations need to be distinguished. First, outliers are observations with a deviating score on the dependent variable. Often (but this does not always have to be the case, as we shall see shortly) outliers are characterized by a high residual value. The residual plots introduced earlier in this chapter can be useful tools to identify outliers.

Second, cases that have extreme scores on the predictor variables are called *leverage points*. Exceptional values on an independent variable can function as a lever, tilting the regression line towards them. The more extreme the *X*-values, the more powerful the lever is. More formally speaking, the degree of leverage of an observation *i* can be defined as the impact that the *X*-values of observation *i* have on the predicted value of that same observation *i*. Leverage values can be calculated as follows. From the previous chapter we know that, using matrix notation, the predicted values of a regression model can be expressed as:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (5.28)$$

where y is an $n \times 1$ vector containing the predicted values, **y** is an $n \times 1$ vector of the observed values for the dependent variable, and **X** is an $n \times p$ matrix of containing the scores on the predictor variables. The impact of the *X*-values on the predicted scores is determined by the so-called hat matrix **H**:

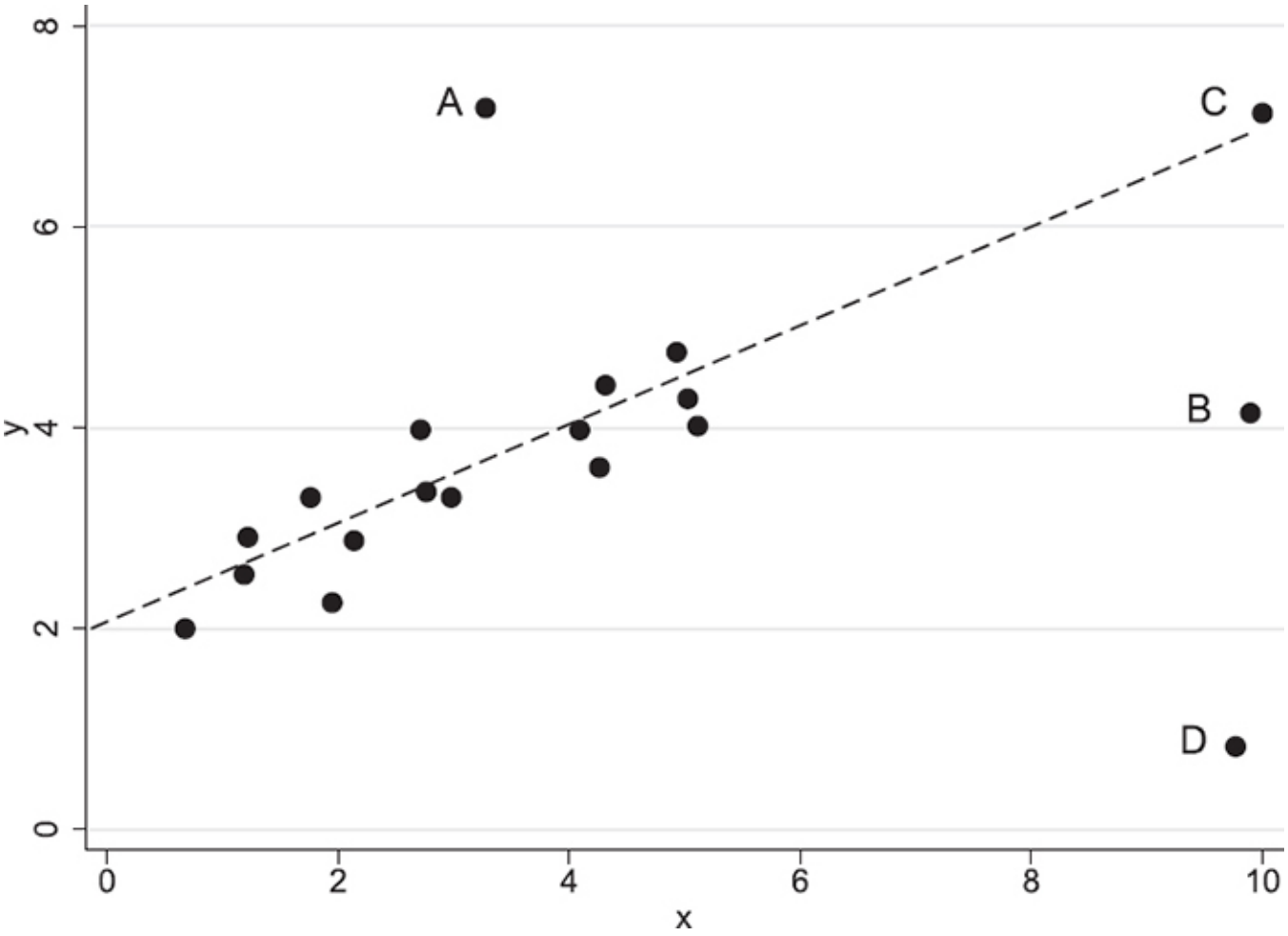
$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \qquad (5.29)$$

which is an n xn matrix. The elements h_{ij} capture the influence that the *X*-values of observation *j* have on the predicted value for observation *i*. In consequence, the diagonal elements (h_{ii}) of the hat matrix are the leverage values as defined above. The higher the value of h_{ii}, the more extreme observation *i* is with respect to the *X*-values. The leverage values h_{ii} have some useful properties. It can be shown that they range between 0 and 1, and that their sum equals the number of regression coefficients to be estimated (p). As a result, the average leverage value equals p/n. A leverage value is considered as large if it is more than twice as large as the average leverage value.

Outliers and leverage points are not necessarily influential cases. Only under certain conditions will they affect the regression estimates. This is illustrated in Figure 5.6, which shows a scatter plot for hypothetical data. Point A has an exceptionally high value on the dependent variables and is thus an outlier. Point B is a leverage point due to its extreme value on the predictor. C and D are at the same time outliers and leverage points. Not all of these points are influential observations. Point A will not influence the regression slope strongly, because there are many other observations with similar values for the predictor. The only influence A might exercise is a slight increase in the intercept of the regression line. Although C is an outlier and has strong leverage, it will not substantially influence the regression line either. The reason is that C is positioned in exactly in the direction of the cloud of dots. B and D, on the other hand, will tilt the regression line severely

in a downward direction. Summarizing, observations are especially influential if they possess leverage and are inconsistent with the regression relation for the other observations.

Figure 5.6 Scatter plot for hypothetical data, showing outliers and leverage points



Diagnostics

Influential observations can be identified by evaluating the impact that single cases have on the regression outcomes. We will discuss three commonly used measures to quantify this impact, namely the DFFITS, Cook's distance and DFBETA. These measures are built on the same general principle. The regression analysis is repeated leaving out a single observation. Subsequently, one assesses how leaving out that observation changes the regression outcomes. This procedure is repeated for every observation in the data set. The three measures differ, however, in the specific outcomes that are evaluated.

The DFFITS (a mnemonic for ‘difference in the fitted value, standardized') measures the change in the predicted value for an observation when the very same observation is left out of the analysis. The DFFITS for observation *i* can be written as

$$DFFITS_i = \frac{\hat{Y}_i - \widehat{Y_{i(i)}}}{\sqrt{MSE_{(i)}h_{ii}}} \quad (5.30)$$

where $\hat{Y}_i$ and $\widehat{Y_{i(i)}}$ are the predicted values for observation *i* for a regression model respectively including and excluding observation *i*, $MSE(i)$ is the mean squared error for the model excluding observation *i*, and h_{ii} is the leverage value of observation *i*. Alternatively, the formula for DFFITS can be written as

$$DFFITS_i = t_i\sqrt{\frac{h_{ii}}{1 - h_{ii}}} \quad (5.31)$$

with t_i equal to the studentized deleted residual, that is, the difference between the observed score Y_i and $\widehat{Y_{i(i)}}$ divided by its estimated standard error. The studentized deleted residual expresses how inconsistent an observation is with the remainder of the cases. Thus, the more inconsistent an observation and the larger its leverage, the larger the DFFITS will be. Observations with a DFFITS larger than $2\sqrt{\frac{p}{(n-p)}}$ are considered influential observations.

While the DFFITS measures the influence of observation *i* on its own predicted value, Cook's distance (D_i) summarizes the impact of observation *i* on predicted values for all other observations:

$$D_i = \frac{\sum_{k=1}^{n}(\hat{Y}_k - \widehat{Y_{k(i)}})^2}{p\ MSE}, \quad (5.32)$$

which is equivalent to

$$D_i = \frac{e_i^2}{p\ MSE}\frac{h_{ii}}{(1 - h_{ii})^2}. \quad (5.33)$$

As with DFFITS, higher residuals and leverage values will lead to a larger Cook's distance. D_i has the advantage, however, that it is less sensitive than DFFITS. Furthermore, strongly influential observations will stand out more clearly since residuals as well as leverage values are squared in the calculation of Cook's distance (Freund and Wilson, 1998, p. 130).

To determine whether observations are influential, the D_i obtained are sometimes compared with the 50th percentile of an *F* distribution with *p* and $n - p$ degrees of freedom (Neter et al., 1996, p. 409). If D_i is larger than this value, observation *i* is considered as influential. Others have argued that $4/n$ should be used as a cut-off point instead (Bollen and Jackman, 1990).

The DFBETAs, finally, quantify how strongly a single observation influences the estimated regression parameters rather than predicted values. For each observation *i* and independent variable *j*, a DFBETA can be calculated as follows:

$$DFBETA_{j(i)} = \frac{b_j - b_{j(i)}}{\sqrt{MSE_{(i)}c_{jj}}}, \qquad (5.34)$$

where b_j is the parameter estimate for independent variable j calculated on the complete data set and $bj(i)$ is the same parameter estimate but then calculated excluding observation i. C_{jj} refers to the jth diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$.

A positive (negative) DFBETA indicates that excluding a particular observation would lead to a decrease (increase) in the respective regression parameters. A DFBETA with an absolute value greater than $2/\sqrt{n}$ is considered large, and points in the direction of an influential observation.

Alternative methods for detecting influential observations are discussed in Freund and Wilson (1998, Chapter 4).

To illustrate the detection of influential observations, we rerun the regression analysis, save the leverage measures and calculate how many cases exceed the critical value (for Cook's distance, the 4/*n* rule is used). Based on Cook's distance and the DFBETAs, 9–11% of the observations can be considered influential. The DFFITS gives a more conservative estimate of 5.9%. Table 5.5 lists the cases with the ten highest values on each of those measures (in the case of the DFBETAs, the highest absolute values/deviations from 0).

Table 5.5 Ten highest values on Cook's distance, DFFITS and DFBET As (Case numbers in square brackets)

Rank	Cook's D		DFFITS		DFBETA age		DFBETA female		DFBETA hincfel		DFBETA eduyrs		DFBETA plcpvcr	
	$Cv^* = 0.0024$		$Cv = 0.0061$		$Cv = 0.0493$		$Cv = 0.0493$		$Cv = 0.0493$		$Cv = 0.0493$		$Cv = 0.0493$	
1	0.0286	[1509]	0.3397	[965]	0.1837	[1535]	0.1271	[1509]	0.2385	[1509]	0.2417	[1504]	0.2843	[121]
2	0.0198	[121]	0.2620	[64]	0.1625	[862]	0.1229	[121]	0.2131	[1119]	0.2294	[1091]	0.2668	[965]
3	0.0191	[965]	0.2391	[1091]	0.1533	[560]	0.1124	[306]	0.1692	[376]	0.2103	[862]	0.2399	[1509]
4	0.0175	[862]	0.2341	[1369]	0.1365	[1426]	0.1045	[965]	0.1582	[8]	0.1963	[1573]	0.2142	[64]
5	0.0116	[1504]	0.2330	[1573]	0.1320	[121]	0.0999	[1119]	0.1441	[306]	0.1698	[756]	0.1783	[1028]
6	0.0114	[64]	0.2177	[1028]	0.1284	[1573]	0.0934	[560]	0.1359	[407]	0.1472	[1509]	0.1594	[87]
7	0.0102	[1119]	0.2007	[927]	0.1172	[1369]	0.0922	[862]	0.1354	[777]	0.1546	[1580]	0.1498	[1270]
8	0.0098	[306]	0.1809	[1581]	0.1151	[149]	0.0903	[376]	0.1282	[1614]	0.1387	[977]	0.1496	[927]
9	0.0095	[1091]	0.1803	[8]	0.1138	[911]	0.0830	[336]	0.1196	[965]	0.1385	[534]	0.1450	[911]
10	0.0095	[560]	0.1753	[185]	0.1124	[650]	0.0812	[1535]	0.1150	[1379]	0.1252	[196]	0.1397	[335]

Associated case numbers in square brackets.
* Cv = critical value

There are some very extreme values, and it is instructive to look at the associated case numbers. Two cases appear most frequently in the top ten lists (1509 and 965). Are there any logical explanations for their influence on the parameter estimates of our regression model? Listing their values on the model variables reveals that both have the maximum distance between trust in the police (our dependent variable) and rating police effectiveness in crime prevention. Since there is a significantly positive relationship between those variables in our model, the extremely negative relationship between them for these observations strongly influences the parameter estimates. Not surprisingly, case 965 and 1509 are both in the top three of highest DFBETAs for the crime prevention predictor.

In our previous diagnostic graphs, one could identify a clear leverage point on educational level. This case,

observation 1091, also appears three times in the list of extreme values and has an especially high DFBETA for education. A list of its values on the model variables shows that this observation concerns a 78-year-old woman shown as having completed no less than 45 years of full-time education!

Remedies

How influential observations should be remedied is, strictly speaking, not a statistical problem, and as such no straightforward statistical solutions exist (Freund and Wilson, 1998, p. 143). Influential observations represent an anomaly in the regression model. The reasons for this anomaly can be manifold. Further investigation of the nature of the influential observation is required before remedying action can be undertaken.

First, influential observations might be caused by errors made during the data collection process. For example, respondents might have provided an incorrect (and implausible) score, or a mistake might have occurred during coding or inputting (this is probably the reason for the exceptional value of observation 1091 on eduyrs). If there is convincing evidence that a data collection error is present, the incorrect value should be corrected, or alternatively the influential observation can be excluded from the analysis.

Second, some observations might behave differently compared to the rest of the data set because they are subject to a factor not accounted for in the model. In this case, influential observations are informative, because they point in the direction of model misspecification. Here, in-depth study of the influential observations is warranted to identify the factor causing the influential observations to behave differently. When successful, the regression model can be respecified by including this factor in the model.

In practice, however, it is often not easy to determine the reasons for the exceptional behavior of some observations. When uncertain about the nature of the influential observations, it is not good practice to simply remove these observations from the data set. Obtaining a better model fit is not a valid argument for modifying the data. In this case, robust regression estimation (a technique that is less sensitive to single observations having exceptional scores) might be considered. A discussion of robust regression estimation can be found in Carroll and Ruppert (1988, Chapter 6).

Assumption 6: Absence of Multicollinearity

What Is It?

In multiple regression analysis, independent variables are not only often related to the dependent variable, but also regularly correlated among themselves. Multivariate regression analysis was specifically designed as a statistical tool to deal with the situation of several correlated variables predicting a single outcome variable. However, if the correlations between the independent variables become too strong, problems can arise during the analysis. Multicollinearity refers to this situation where (a set of) predictor variables show very strong intercorrelations. We speak of perfect multicollinearity when one independent variable can be perfectly

predicted by the other independent variables.

Consequences

A first consequence of multicollinearity is that the interpretability of the results of the regression analysis is made more difficult. As a multivariate analytical tool, the aim of regression analysis is to disentangle the effects of several predictors and to estimate net effects of one independent variable, keeping other predictors constant. Yet when strong multicollinearity is present, this task becomes meaningless. If predictors are almost perfectly intertwined, attempts to disentangle them are not fruitful. And the idea of increasing one predictor while keeping all other variables constant is not meaningful if predictors covary almost perfectly. Two strongly correlated predictors might turn out to have very small and even insignificant effects, even if separately they have strong predictive power. In sum, multicollinear data makes regression coefficients hard to interpret.

The statistical consequences of multicollinearity can be illustrated by looking at the formula for the point estimators for the regression coefficients:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \qquad (5.35)$$

In this formula, the inverse of **X'X** is used. This inverse can be calculated by dividing the cofactor matrix of **X'X** by the determinant of **X'X**:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{(\mathbf{X}'\mathbf{X})_{\text{cof}}}{|\mathbf{X}'\mathbf{X}|}. \qquad (5.36)$$

If a perfect linear dependency between two or more predictors exists, then matrix **X** is not of full rank. As a result, (**X'X**) will also not be of full rank, and the determinant |**X'X**| will equal 0. The consequence is that the inverse of (**X'X**) is not defined, and that the regression coefficients cannot be estimated.

In practice, however, perfect multicollinearity is rather exceptional. But also when the intercorrelations between predictors are very strong instead of perfect, statistical problems are encountered. In this case, the determinant |**X'X**| will not equal zero, but be very close to zero. Because this determinant figures in the denominator, $(\mathbf{X}'\mathbf{X})^{-1}$ will be inflated. In the first place, this results in very unstable estimates for the regression coefficients. After all, the vector of regression coefficients **b** is the product of $(\mathbf{X}'\mathbf{X})^{-1}$ and **X'y**. Small changes in **X'y** will lead to substantial differences in the regression coefficients. Furthermore, the inverse of **X'X** is used in the estimation of the variance–covariance matrix of the regression coefficients (cf. Chapter 4 of this volume):

$$\sigma_b^2 = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad (5.37)$$

Multicollinearity thus also increases the standard errors of regression coefficients, and renders it more difficult to find statistically significant effects.

The consequences of multicollinearity are thus potentially very severe. In social science research (where perfect relationships between variables are exceptional), however, this phenomenon mainly occurs in very specific situations. There is a risk of multicollinearity, for example, when interaction effects or higher-order

polynomial terms are introduced in the model (in these cases, centering the predictors prior to calculating interactions or polynomial terms can remedy the problem; cf. Brambor et al., 2006).

Diagnostics

The presence of multicollinearity can be tested formally by making use of the so-called tolerance and *variance inflation factor* statistics. Both measures quantify the extent to which a certain predictor *j* depends on the set of other predictors.

The tolerance of a predictor *j* equals 1 minus the proportion of explained variance of a regression model explaining predictor *j* by means of the other independent variables:

$$TOL_j = 1 - R_j^2. \qquad (5.38)$$

In other words, the tolerance expresses the amount of unique variance in a predictor. Tolerance values range from 0 to 1. Small tolerance values are indicative of multicollinearity, as they imply that a predictor depends strongly on the other independent variables. As a rule of thumb, tolerance values smaller than 0.1 are considered problematic.

The variance inflation factor is defined as the inverse of the tolerance:

$$VIF_j = \frac{1}{TOL_j} = \frac{1}{1 - R_j^2}. \qquad (5.39)$$

Intuitively, *VIFj* can be interpreted as the factor by which the variance of independent variable *j* increases due to the intercorrelations with the other variables. A variance inflation factor of 2, for example, means that the variance of *Xj* in the multivariate model has doubled compared to a model where *Xj* would be the only predictor variable, thereby leading to less stable estimates and larger standard errors. The variance inflation factor ranges between 1 and +∞. High values indicate that the respective predictor depends more strongly on the other independent variables. Variance inflation factor values larger than 10 indicate that potentially harmful multicollinearity is present.

Table 5.6 shows the variance inflation factors for the regression model explaining trust in the police. In this example, all VIFs are close to 1, and they come nowhere close to the cut-off value of 10. We conclude that no multicollinearity is present.

Table 5.6 Variance inflation factors

Variable	VIF
Age	1.07
Female	1.06
Hincfel	1.05
Eduyrs	1.01

Plcpvcr	1.01

Remedies

The most straightforward remedy for multicollinearity is to remove one or more predictors with strong intercorrelations. Alternatively, data reduction techniques, such as principal components or factor analysis, can be used to summarize the information contained by a set of interrelated predictors by a more limited number of factors.

Sometimes, removing variables or data reduction is not an option for theoretical reasons. In that case, ridge regression can be applied. The idea behind ridge regression is that the linear dependency between the columns of matrix **X'X** can be reduced by adding a constant value to the diagonal elements of this matrix. As a result, the multicollinearity problem will be attenuated, but small biases in the parameter estimates can be introduced. Because multicollinearity is encountered only rarely in applied social research, we do not discuss ridge regression in detail here, but refer to Neter et al. (1996, pp. 394ff.).

Concluding Remarks

This chapter has provided an overview of six often-mentioned assumptions underlying OLS regression. Three of these conditions concern the error terms of the regression equation: error terms are assumed to be homoscedastic, independent and normally distributed. Other assumptions regard the functional form of relations between independent and dependent variables (linearity), interrelations among predictors (absence of multicollinearity) or the position of individual data points (absence of influential observations).

If these assumptions are not fulfilled, regression results can be distorted and conclusions might be misleading. The nature and severity of the consequences of violations, however, vary greatly from one assumption to the other. Generally speaking, regression analysis has two main functions. First, it is employed as an analytical tool to describe the structure of the observed data. This first purpose requires accurate point estimates of regression coefficients. Violations of the assumption of linearity and the presence of influential observations, however, can cause bias in the regression parameters. Furthermore, multicollinearity among the predictors can cause a large amount of unreliability in regression estimates. As such, violations of these three assumptions can pose a threat to the descriptive function of regression analysis. Second, regression is also often used to make statistical inferences, and generalize findings to a wider population. For this second purpose, not only point estimates but also estimated standard errors of regression parameters are of great importance. Heteroscedasticity, non-normality and non-independence of residuals each potentially create bias in the estimates of standard errors. Additionally, multicollinearity can lead to unreliable estimates of standard errors. In consequence, violations of these assumptions can hamper statistical inference.

Yet, it has to be mentioned that, from the perspective of applied researchers, not all assumptions are equally

important. When the size of the sample analyzed is sufficiently large, for example, deviations from normality usually do not have harmful consequences. Also violations of the assumption of homoscedasticity are in many cases relatively minor (Gelman and Hill, 2007, p. 46).

The list of assumptions presented in this chapter is not exhaustive, however, and compliance with these six assumptions is not a sufficient condition for drawing valid and reliable conclusions from the regression model. As is the case in any statistical model, the strength of conclusions crucially depends on the measurement quality of the variables (Carmines and Zeller, 1979). As such, regression analysis assumes that measurements are valid, that predictor variables are not contaminated by random measurement error,[3] and that the dependent variable is measured on an interval scale. Furthermore, regression assumes that no model misspecifications are present and that no important causal factors are left out of the model (so-called omitted variable bias) – refer to Berry (1993) for a more detailed treatment of these issues.

Finally, it has to be repeated that this chapter reviews assumptions for the OLS regression model specifically. Assumptions for other regression models are discussed elsewhere – see Harrell (2001) for logistic regression and survival analysis; and Snijders and Bosker (2012) for multilevel models.

Notes

1 This chapter focuses on the assumptions of OLS regression, and does not deal with the assumptions underlying other regression models (e.g. logistic regression and multilevel modelling). A thorough discussion of assumptions for these models can be found elsewhere – see Harrell (2001) for logistic regression and Snijders and Bosker (2012) for multilevel models.
2 Various alternatives to perform the lack-of-fit test are possible, such as an ANOVA decomposition of error or a likelihood ratio test comparing a regression where predictors are specified as categorical and one where predictors are specified as continuous. These alternatives lead to identical results.
3 Random measurement error in the dependent variable is less problematic, since this source of error is accommodated in the model by means of the residual term.

References

Allen, M. P. J. (**1997**). *Understanding Regression Analysis*. New York: Plenum Press.

Berry, W. D. (**1993**). *Understanding Regression Assumptions, volume 07-092 of Quantitative Applications in the Social Sciences*. Newbury Park, CA: Sage.

Bollen, K. A. and **Jackman, R. W.** (**1990**). Regression diagnostics: An expository treatment of outliers and influential cases. In Edited by: **J.Fox** and **J. S.Long** (Eds), *Modern Methods of Data Analysis*. Newbury Park, CA: Sage.

Brambor, T., **Clark, W. R.** and **Golder, M.** (**2006**). Understanding interaction models: Improving empirical analyses. *Political Analysis*, 14(1), 63–82. http://dx.doi.org/10.1093/pan/mpi014

Carmines, E. G. and **Zeller, R. A.** (**1979**). *Reliability and Validity Assessment*. Newbury Park, CA: Sage.

Carroll, R. J. and **Ruppert, D.** (**1988**). *Transformation and Weighting in Regression*. New York: Chapman & Hall. http://dx.doi.org/10.1007/978-1-4899-2873-3

DeMaris, A. (**2004**). *Regression with Social Data: Modeling Continuous and Limited Response Variables*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/0471677566

Fox, J. (**2008**). *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks, CA: Sage.

Freund, R. J. and **Wilson, W. J.** (**1998**). *Regression Analysis: Statistical Modeling of a Response Variable*. San Diego, CA: Academic Press.

Gelman, A. and **Hill, J.** (**2007**). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Harrell, F. E. (**2001**). *Regression Modelling Strategies*. New York: Springer. http://dx.doi.org/10.1007/978-1-4757-3462-1

Krzanowski, W. (**1998**). *An Introduction to Statistical Modelling*. London: Arnold.

Lumley, T., **Diehr, P.**, **Emerson, S.** and **Chen, L.** (**2002**). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23, 151–169. http://dx.doi.org/10.1146/annurev.publhealth.23.100901.140546

Neter, J., **Kutner, M. H.**, **Nachtsheim, C. J.** and **Wasserman, W.** (**1996**). *Applied Linear Statistical Models*. Chicago: Irwin.

Panik, M. (**2009**). *Regression Modeling. Methods, Theory and Computation with SAS*. Boca Raton, FL: Taylor & Francis. http://dx.doi.org/10.1201/9781420091984

Razali, N. and **Wah, Y. B.** (**2011**). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33.

Scariano, S. M. and **Davenport, J. M.** (**1987**). The effects of violations of independence assumptions in the one-way anova. *American Statistician*, 41(2), 123–129.

Shapiro, S. S. and **Wilk, M. B.** (**1965**). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591–611. http://dx.doi.org/10.1093/biomet/52.3-4.591

Snijders, T. A. B. and **Bosker, R. J.** (**2012**). *An Introduction to Basic and Advanced Multilevel Modeling*, 2nd

edn. London: Sage.

White, H. (**1980**). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 48(4), 817–838. http://dx.doi.org/10.2307/1912934

http://dx.doi.org/10.4135/9781446288146.n5