Chapter 7
Statistical Methods: Regression Analysis

Bhimasankaram Pochiraju and Hema Sri Sai Kollipara

1 Introduction

Regression analysis is arguably one of the most commonly used and misused statistical techniques in business and other disciplines. In this chapter, we systematically develop a linear regression modeling of data. Chapter 6 on basic inference is the only prerequisite for this chapter. We start with a few motivating examples in Sect. 2. Section 3 deals with the methods and diagnostics for linear regression. Section 3.1 is a discussion on what is regression and linear regression, in particular, and why it is important. In Sect. 3.2, we elaborate on the descriptive statistics and the basic exploratory analysis for a data set. We are now ready to describe the linear regression model and the assumptions made to get good estimates and tests related to the parameters in the model (Sect. 3.3). Sections 3.4 and 3.5 are devoted to the development of basic inference and interpretations of the regression with single and multiple regressors. In Sect. 3.6, we take the help of the famous Anscombe (1973) data sets to demonstrate the need for further analysis. In Sect. 3.7, we develop the basic building blocks to be used in constructing the diagnostics. In Sect. 3.8, we use various residual plots to check whether there are basic departures from the assumptions and to see if some transformations on the regressors are warranted.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-68837-4_7) contains supplementary material, which is available to authorized users.

B. Pochiraju
Applied Statistics and Computing Lab, Indian School of Business, Hyderabad, Telangana, India

H. S. S. Kollipara (✉)
Indian School of Business, Hyderabad, Telangana, India
e-mail: hemasri.kollipara@gmail.com

B. Pochiraju, S. Seshadri (eds.), *Essentials of Business Analytics*, International Series in Operations Research & Management Science 264,
https://doi.org/10.1007/978-3-319-68837-4_7

Suppose we have developed a linear regression model using some regressors. We find that we have data on one more possible regressor. Should we bring in this variable as an additional regressor, given that the other regressors are already included? This is what is explored through the added variable plot in Sect. 3.9.

In Sect. 3.10, we develop deletion diagnostics to examine whether the presence and absence of a few observations can make a large difference to the quantities of interest in regression, such as regression coefficient estimates, standard errors, and fit of some observations. Approximate linear relationships among regressors are called collinear relationships among regressors. Such collinear relationships can cause insignificance of important regressors, wrong signs for regression coefficient estimates, and instability of the estimates to slight changes in the data. Section 3.11 is devoted to the detection of collinearity and correction through remedial measures such as stepwise regression, best subset regression, ridge regression, and lasso regression. There we note that subset selection deserves special interest. How do we represent categorical regressors? This leads us to the study of dummy variables, which is done in Sect. 3.12. We also consider the case of several categories and ordinal categories. We mention the use of interactions to identify certain effects. Section 3.13 deals with the transformation of response variables and a specific famous method known as Box–Cox transformation. One of the assumptions in linear regression model estimation through least squares is that the errors have equal variance. What if this assumption is violated? What are the ill-effects? How does one detect this violation which is often called heteroscedasticity? Once detected, what remedial measures are available? These questions are dealt with in Sect. 3.14. Another assumption in the estimation of the linear regression model by the method of least squares is that of independence of the errors. We do not deal with this problem in this chapter as this is addressed in greater detail in the Forecasting Analytics chapter (Chap. 12). Section 3.15 deals with the validation of a developed linear regression model. Section 3.16 provides broad guidelines useful for performing linear regression modeling. Finally, Sect. 3.17 addresses some FAQs (frequently asked questions). R codes are provided at the end.

Chatterjee and Hadi (2012) and Draper and Smith (1998) are very good references for linear regression modeling.

2 Motivating Examples

2.1 Adipose Tissue Problem

We all have a body fat called adipose tissue. If abdominal adipose tissue area (*AT*) is large, it is a potential risk factor for cardiovascular diseases (Despres et al. 1991). The accurate way of determining *AT* is computed tomography (CT scan). There are three issues with CT scan: (1) it involves irradiation, which in itself can be harmful to the body; (2) it is expensive; and (3) good CT equipment are not available in smaller towns, which may result in a grossly inaccurate measurement of the area. Is there a way to predict *AT* using one or more anthropological

measurements? Despres et al. (1991) surmised that obesity at waist is possibly a good indicator of large *AT*. So they thought of using waist circumference (*WC*) to predict *AT*. Notice that one requires only a measuring tape for measuring WC. This does not have any of the issues mentioned above for the CT scan. In order to examine their surmise, they got a random sample of 109 healthy-looking adult males and measured WC in cm and *AT* in cm^2 using a measuring tape and CT scan, respectively, for each. The data are available on our data website.

Can we find a suitable formula for predicting *AT* of an individual using their WC? How reliable is this prediction? For which group of individuals is this prediction valid? Is there one formula which is the best (in some acceptable sense)? Are there competing formulae, and if so, how to choose among them?

The dataset "wc_at.csv" is available on the book's website.

2.2 Newspaper Problem

Mr. Warner Sr. owns a newspaper publishing house which publishes the daily newspaper *Newsexpress*, having an average daily circulation of 500,000 copies. His son, Mr. Warner Jr. came up with an idea of publishing a special Sunday edition of *Newsexpress*. The father is somewhat conservative and said that they can do so if it is almost certain that the average Sunday circulation (circulation of the Sunday edition) is at least 600,000.

Mr. Warner Jr. has a friend Ms. Janelia, who is an analytics expert whom he approached and expressed his problem. He wanted a fairly quick solution. Ms. Janelia said that one quick way to examine this is to look at data on other newspapers in that locality which have both daily edition and Sunday edition. "Based on these data," said Ms. Janelia, "we can fairly accurately determine the lower bound for the circulation of your proposed Sunday edition." Ms. Janelia exclaimed, "However, there is no way to pronounce a meaningful lower bound with certainty."

What does Ms. Janelia propose to do in order to get an approximate lower bound to the Sunday circulation based on the daily circulation? Are there any assumptions that she makes? Is it possible to check them?

2.3 Gasoline Consumption

One of the important considerations both for the customer and manufacturer of vehicles is the average mileage (miles per gallon of the fuel) that it gives. How does one predict the mileage? Horsepower, top speed, age of the vehicle, volume, and percentage of freeway running are some of the factors that influence mileage. We have data on MPG (miles per gallon), HP (horsepower), and VOL (volume of

cab-space in cubic feet) for 81 vehicles. Do HP and VOL (often called explanatory variables or regressors) adequately explain the variation in MPG? Does VOL have explaining capacity of variation in MPG over and above HP? If so, for a fixed HP, what would be the impact on the MPG if the VOL is decreased by 50 cubic feet? Are some other important explanatory variables correlated with HP and VOL missing? Once we have HP as an explanatory variable, is it really necessary to have VOL also as another explanatory variable?

The sample dataset[1] was inspired by an example in the book *Basic Econometrics* by Gujarati and Sangeetha. The dataset "cars.csv" is available on the book's website.

2.4 Wage Balance Problem

Gender discrimination with respect to wages is a hotly debated topic. It is hypothesized that men get higher wages than women with the same characteristics, such as educational qualification and age.

We have data on wage, age, and years of education on 100 men and 100 women with comparable distributions of age and years of education. Is it possible to find a reasonable formula to predict wage based on age, years of education, and gender? Once such a formula is found, one can try to examine the hypothesis mentioned above. If it is found that, indeed, there is a gender discrimination, it may be also of interest to examine whether women catch up with men with an increase in educational qualification. After accounting for gender difference and age, is it worthwhile to have higher educational qualification to get a higher wage? Can one quantify result in such a gain?

The dataset "wage.csv" is available on the book's website.

2.5 Medicinal Value in a Leaf

The leaf of a certain species of trees is known to be of medicinal value proportional to its surface area. The leaf is of irregular shape and hence it is cumbersome to determine the area explicitly. One scientist has thought of two measurements (which are fairly easy to obtain), the length and breadth of the leaf which are defined as follows: the length is the distance between the two farthest points in the leaf and the breadth is the distance between the two farthest points in the leaf perpendicular to the direction of the length. The scientist obtained the length, breadth, and area (area measured in the laborious way of tracing the leaf on a graph paper and counting the squares in the traced diagram) on 100 randomly selected leaves. Is it possible to find

[1] Original datasource is US Environmental Pollution Agency (1991), Report EPA/AA/CTAB/91-02 and referred to in the book "Basic Econometrics" by Gujarati and Sangeetha.

an approximate formula for the area of the leaf based on its length and breadth which are relatively easy to measure? The dataset "leaf.csv" is available on the book's website.

3 Methods of Regression

3.1 What Is Regression?

Let us consider the examples in Sect. 2. We have to find an approximate formula—for AT in terms of WC in Example 2.1, for the circulation of Sunday edition in terms of the circulation of daily edition in Example 2.2, for MPG in terms of HP, VOL, and WT in Example 2.3, for wage in terms of age, years of education, and gender in Example 2.4, and for the area of the leaf in terms of its length and width in Example 2.5. We call the variable of interest for which we want to get an approximate formula as the *response variable*. In the literature, the response variable is synonymously referred to as the *regressand*, and the *dependent variable* also. The variables used to predict the response variable are called *regressors*, *explanatory variables*, *independent variables*, or *covariates*. In Example 2.4, wage is the response variable and age, years of education, and gender are the regressors.

In each of these examples, even if the data are available on all the units in the population, the formula cannot be exact. For example, AT is not completely determined by WC. It also depends on gender, weight, triceps, etc. (Brundavani et al. 2006). Similarly, in other examples it is easy to notice that regressors do not completely determine the response variable. There are omitted variables on which we do not have data. We have a further limitation that the data are available on a sample and hence we can only estimate the prediction formula. Thus, there are two stages in arriving at a formula: (a) to postulate a functional form for the approximation which involves some parameters and (b) to estimate the parameters in the postulated functional form based on the sample data.

If we denote the response variable by Y, the regressors by $X_1, \ldots, X_k$, the parameters by $\theta_1, \ldots, \theta_r$, and the combined unobserved variables, called the error, ε, then we attempt to estimate an equation:

$$Y = f\left(X_1, \ldots, X_k, \theta_1, \ldots, \theta_r, \varepsilon\right). \tag{7.1}$$

In Example 2.2, we may postulate the formula as

$$Sunday\ circulation = \alpha + \beta\ daily\ \ circulation + \varepsilon. \tag{7.2}$$

Here the functional form f is linear in the regressor, namely, daily circulation. It is also linear in the parameters, α and β. The error ε has also come into the equation as an additive term.

In Example 2.4, notice that a natural functional form of area in terms of the length and breadth is multiplicative. Thus we may postulate

$$area = \alpha.length^{\beta_1}.breadth^{\beta_2}.\varepsilon \tag{7.3}$$

Here the functional form is multiplicative in powers of length and breadth and the error. It is not linear in the parameters either.

Alternatively, one may postulate

$$area = \alpha.length^{\beta_1}.breadth^{\beta_2} + \varepsilon \tag{7.4}$$

In this specification, the functional form is multiplicative in powers of length and breadth but additive in the error. It is not linear in the parameters either.

How does one know which of the two specifications (7.3) and (7.4) is appropriate? Or is it that neither of them is appropriate? There are ways to examine this based on the data. We shall deal with this in detail subsequently.

3.2 What Is Linear Regression?

If the functional form, *f*, in Eq. (7.1), is linear in the parameters, then the regression is called linear regression. As noted earlier, (7.2) is a linear regression equation. What about Eq. (7.3)? As already noted, this is not a linear regression equation. However, if we make a log transformation on both sides, we get

$$\log(area) = \log\alpha + \beta_1 \log(length) + \beta_2 \log(breadth) + \log\varepsilon \tag{7.5}$$

which is linear in the parameters: $\log\alpha$, β_1, *and* β_2. However, this model is not linear in length and breadth.

Such a model is called an *intrinsically linear regression model.*

However, we cannot find any transformation of the model in (7.4) yielding a linear regression model. Such models are called *intrinsically nonlinear regression models*.

3.3 Why Regression and Why Linear Regression?

Regression is performed for one or more of the following reasons:

1. To predict the response variable for a new case based on the data on the regressors for that case.

2. To study the impact of one regressor on the response variable keeping other regressors fixed. For example, one may be interested in the impact of one additional year of education on the average wage for a person aged 40 years.
3. To verify whether the data support certain beliefs (hypotheses)—for example, whether $\beta_1 = \beta_2 = 1$ in the Eq. (7.3) which, if upheld, would mean that the leaf is more or less rectangular.
4. To use as an intermediate result in further analysis.
5. To calibrate an instrument.

Linear regression has become popular for the following reasons:

1. The methodology for linear regression is easily understood, as we shall see in the following sections.
2. If the response variable and the regressors have a joint normal distribution, then the regression (as we shall identify in Sect. 3.5, regression is the expectation of the response variable conditional on the regressors) is a linear function of the regressors.
3. Even though the model is not linear in the regressors, sometimes suitable transformations on the regressors or response variable or both may lead to a linear regression model.
4. The regression may not be a linear function in general, but a linear function may be a good approximation in a small focused strip of the regressor surface.
5. The methodology developed for the linear regression may also act as a good first approximation for the methodology for a nonlinear model.

We shall illustrate each of these as we go along.

3.4 Basic Descriptive Statistics and Box Plots

Analysis of the data starts with the basic descriptive summary of each of the variables in the data (the Appendix on Probability and Statistics provides the background for the following discussion). The descriptive summary helps in understanding the basic features of a variable such as the central tendency, the variation, and a broad empirical distribution. More precisely, the basic summary includes the minimum, the maximum, the first and third quartiles, the median, the mean, and the standard deviation. The minimum and the maximum give us the range. The mean and the median are measures of central tendency. The range, the standard deviation, and the interquartile range are measures of dispersion. The box plot depicting the five-point summary, namely, the minimum, the first quartile, the median, the third quartile, and the maximum, gives us an idea of the empirical distribution. We give below these measures for the (WC, AT) data.

From Table 7.1 and Fig. 7.1, it is clear that the distribution of WC is fairly symmetric, about 91 cm, and the distribution of AT is skewed to the right. We shall see later how this information is useful.

Table 7.1 Descriptive statistics for (WC, AT) data

Variable	min	1st Qu	Median	Mean	3rd Qu	Max	Std. dev.
WC (in cm)	63.5	80.0	90.8	91.9	104.0	121.0	13.55912
AT (in sq. cm)	11.44	50.88	96.54	101.89	137.00	253.00	57.29476

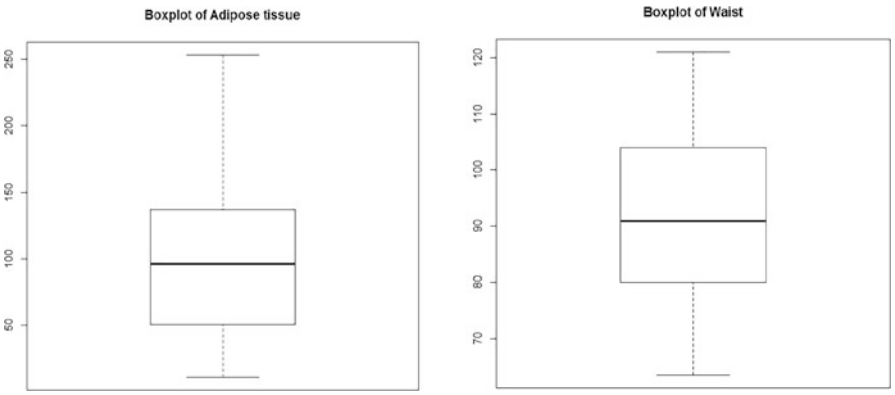


Fig. 7.1 Box plots of adipose tissue area and waist circumference

3.5 Linear Regression Model and Assumptions

The linear regression model can be written as follows:

$$\left.\begin{array}{l} Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \\ E\left(\varepsilon | X_1, \ldots, X_k\right) = 0, \\ Var\left(\varepsilon | X_1, \ldots, X_k\right) = \sigma^2 \left(> 0\right) \end{array}\right\} \tag{7.6}$$

Note on notation: The notation E(Y|X) stands for the expected value of Y given the value of X. It is computed using P(Y|X), which stands for the conditional probability given X. For example, we are given the following information: when X = 1, Y is normally distributed with mean 4 and standard deviation of 1; whereas, when X = 2, Y is normally distributed with mean 5 and standard deviation of 1.1. E(Y|X = 1) = 4 and E(Y|X = 2) = 5. Similarly, the notation Var(Y|X) is interpreted as the variance of Y given X. In this case Var(Y|X = 1) = 1 and Var(Y|X = 2) = 1.21.

The objective is to draw inferences (estimation and testing) related to the parameters $\beta_0, \ldots, \beta_k, \sigma^2$ in the model (7.6) based on the data $(y_i, x_{i1}, \ldots, x_{ik}), i = 1, \ldots, N$ on a sample with N observations from $(Y, X_1, \ldots, X_k)$. Note that Y is a column vector of size N × 1. The transpose of a vector Y (or matrix M) is written as Y^t (M^t), the transpose of a column vector is a row vector and vice versa.

We now write down the observational equations as

$$\left.\begin{array}{l} y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ . \\ . \\ . \\ y_N = \beta_0 + \beta_1 x_{N1} + \cdots + \beta_k x_{Nk} + \varepsilon_N \end{array}\right\} \tag{7.7}$$

.

The above model can be written compactly as

$$\left.\begin{array}{l} Y = Z\beta + \varepsilon \\ where \\ Y = (y_1, \ldots, y_N)^t, X = \left((x_{ij})\right), Z = (1 : X), \varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^t \end{array}\right\} \tag{7.8}$$

In (7.8), by $X = ((x_{ij}))$, we mean that X is a matrix, the element in the junction of the i^{th} row and j^{th} column of which is x_{ij}. The size of X is $N \times k$. In the matrix Z, 1 denotes the column vector, each component of which is the number 1. The matrix 1 is appended as the first column and the rest of the columns are taken from X—this is the notation (1:X). The matrices X and Z are of orders $N \times k$ and $N \times (k+1)$, respectively.

We make the following assumption regarding the errors $\varepsilon_1, \ldots, \varepsilon_N$:

$$\varepsilon \mid X \sim N_N\left(0, \sigma^2 I\right) \tag{7.9}$$

that is, the distribution of the errors given X is an N-dimensional normal distribution with mean zero and covariance matrix equal to σ^2 times the identity matrix. In (7.9), I denotes the identity matrix of order $N \times N$. The identity matrix comes about because the errors are independent of one another, therefore, covariance of one error with another error equals zero.

From (7.8) and (7.9), we have

$$Y \mid X \sim N_N\left(Z\beta, \sigma^2 I\right).$$

In other words, the regression model Zβ represents the mean value of Y given X. The errors are around the mean values.

What does the model (7.7) (or equivalently 7.8) together with the assumption (7.9) mean? It translates into the following:

1. The model is linear in the parameters $\beta_0, \ldots, \beta_k$(L).
2. Errors conditional on the data on the regressors are independent (I).
3. Errors conditional on the data on the regressors have a joint normal distribution (N).

4. The variance of each of the errors conditional on the data on the regressors is σ^2(E).
5. Each of the errors conditional on the data on the regressors has 0 mean.

In (4) above, E stands for equal variance. (5) above is usually called the exogeneity condition. This actually implies that the observed covariates are uncorrelated with the unobserved covariates. The first four assumptions can be remembered through an acronym: LINE.

Why are we talking about the distribution of $\varepsilon_i \mid X$? Is it not a single number? Let us consider the adipose tissue example. We commonly notice that different people having the same waist circumference do not necessarily have the same adipose tissue area. Thus there is a distribution of the adipose tissue area for people with a waist circumference of 70 cm. Likewise in the wage example, people with the same age and education level need not get exactly the same wage.

The above assumptions will be used in drawing the inference on the parameters. However, we have to check whether the data on hand support the above assumptions. How does one draw the inferences and how does one check for the validity of the assumptions? A good part of this chapter will be devoted to this and the interpretations.

3.6 Single Regressor Case

Let us consider Examples 2.1 and 2.2. Each of them has one regressor, namely, WC in Example 2.1 and daily circulation in Example 2.2. Thus, we have bivariate data in each of these examples. What type of relationship does the response variable have with the regressor? We have seen in the previous chapter that covariance or correlation coefficient between the variables is one measure of the relationship. We shall explore this later where we shall examine the interpretation of the correlation coefficient. But we clearly understand that it is just one number indicating the relationship. However, we do note that each individual (WC, AT) is an ordered pair and can be plotted as a point in the plane. This plot in the plane with WC as the X-axis and AT as the Y-axis is called the scatterplot of the data. We plot with the response variable on the Y-axis and the regressor on the X-axis. For the (WC, AT) data the plot is given below:

What do we notice from this plot?

1. The adipose tissue area is by and large increasing with increase in waist circumference.
2. The variation in the adipose tissue area is also increasing with increase in waist circumference.

The correlation coefficient for this data is approximately 0.82. This tells us the same thing as (1) above. It also tells that the strength of (linear) relationship between the two variables is strong, which prompts us to fit a straight line to the data. But by looking at the plot, we see that a straight line does not do justice for large values of

waist circumference as they are highly dispersed. (More details on this later.) So the first lesson to be learnt is: *If you have a single regressor, first look at the scatterplot.* This will give you an idea of the form of relationship between the response variable and the regressor. If the graph suggests a linear relationship, one can then check the correlation coefficient to assess the strength of the linear relationship between the response variable and the regressor.

We have the following linear regression model for the adipose tissue problem:

$$\text{Model}: \quad AT = \beta_0 + \beta_1 WC + \varepsilon \tag{7.10}$$

$$\text{Data}: \quad (AT_i, WC_i), i = 1, ...109 \tag{7.11}$$

$$\text{Model adapted to data}: \quad AT_i = \beta_0 + \beta_1 WC_i + \varepsilon_i, i = 1, \ldots, 109 \tag{7.12}$$

Assumptions: $\varepsilon_i \mid WC_i, i = 1, \ldots, 109$ are independently and identically distributed as normal with mean 0 and variance σ^2 often written in brief as

$$\varepsilon_i \mid WC_i, i = 1, \ldots, 109 \quad \text{are} \quad iid \quad N\left(0, \sigma^2\right) \quad \text{variables.} \tag{7.13}$$

Model described by (7.12) and (7.13) is a special case of the model described by (7.7) and (7.9) where k = 1 (For a single regressor case, k = 1, the number of regressors.) and N = 109.

A linear regression model with one regressor is often referred to as a *Simple Linear Regression* model.

Estimation of Parameters

From the assumptions it is clear that all the errors (conditional on the Waist Circumference values) are of equal importance.

If we want to fit a straight line for the scatterplot in Fig. 7.2, we look for that line for which some reasonable measure of the magnitude of the errors is small. A straight line is completely determined by its intercept and slope, which we denote by β_0 and β_1, respectively. With this straight line approximation, from (7.12), the error for the i^{th} observation is $\varepsilon_i = AT_i - \beta_0 - \beta_1 WC_i, i = 1, \ldots, 109$.

A commonly used measure for the magnitude of the errors is the sum of their squares, namely, $\sum_{i=1}^{109} \varepsilon_i^2$. So if we want this measure of the magnitude of the errors to be small, we should pick up the values of β_0 and β_1, which will minimize $\sum_{i=1}^{109} \varepsilon_i^2$. This method is called the *Method of Least Squares*. This is achieved by solving the equations (often called *normal equations*):

$$\left.\begin{array}{l} \sum_{i=1}^{109} AT_i = 109\beta_0 + \beta_1 \sum_{i=1}^{109} WC_i \\ \sum_{i=1}^{109} AT_i WC_i = \beta_0 \sum_{i=1}^{109} WC_i + \beta_1 \sum_{i=1}^{109} WC_i^2 \end{array}\right\} \tag{7.14}$$

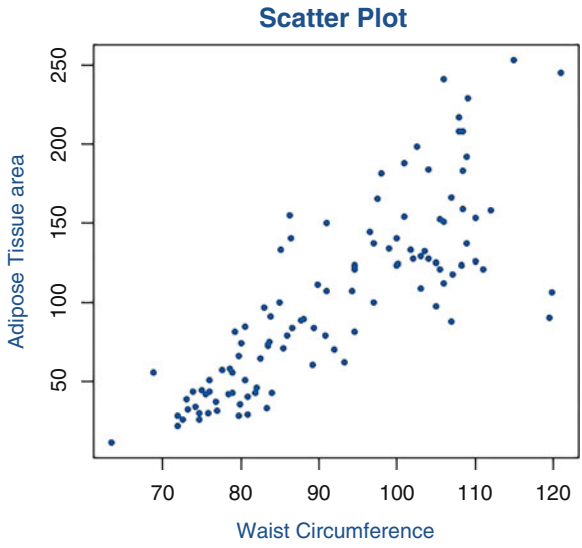


Fig. 7.2 Lo and behold! A picture is worth a thousand words

Solving the two equations in (7.14) simultaneously, we get the estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$ of β_0,β_1, respectively, as

$$\left.\begin{aligned}\widehat{\beta}_1 &= \frac{\text{cov}(AT,WC)}{V(WC)} = \frac{\sum_{i=1}^{109} AT_i WC_i - \sum_{i=1}^{109} AT_i \sum_{i=1}^{109} WC_i/109}{\sum_{i=1}^{109} WC_i^2 - \left(\sum_{i=1}^{109} WC_i\right)^2/109} \\ \widehat{\beta}_0 &= \overline{AT} - \widehat{\beta}_1 \overline{WC}\end{aligned}\right\} \tag{7.15}$$

(Let $v_1, \ldots, v_r$ be a sample from a variable V. By $\overline{V}$, we mean the average of the sample.)

Thus the estimated regression line, often called the *fitted model*, is

$$\widehat{AT} = \widehat{\beta}_0 + \widehat{\beta}_1 WC. \tag{7.16}$$

The predicted value (often called the *fitted value*) of the i^{th} observation is given by

$$\widehat{AT_i} = \widehat{\beta}_0 + \widehat{\beta}_1 WC_i. \tag{7.17}$$

Notice that this is the part of the adipose tissue area for the i^{th} observation explained by our fitted model.

The part of the adipose tissue area of the i^{th} observation, not explained by our fitted model, is called the *Pearson residual*, henceforth referred to as *residual* corresponding to the i^{th} observation, and is given by

$$e_i = AT_i - \widehat{AT_i}. \tag{7.18}$$

Thus, $AT_i = \widehat{AT_i} + e_i = \widehat{\beta}_0 + \widehat{\beta}_1 WC_i + e_i$. Compare this with (7.12). We notice that $\widehat{\beta}_0, \widehat{\beta}_1, e_i$ are the sample analogues of β_0,β_1,ε_i respectively. We know that the errors, ε_i are unobservable. Therefore, we use their sample representatives e_i to check the assumptions (7.9) on the errors.

The sum of squares of residuals, $R_0^2 = \sum_{i=1}^{109} e_i^2$ is the part of the variation in the adipose tissue area that is not explained by the fitted model (7.16) estimated by the method of least squares. The estimator of σ^2 is obtained as

$$\widehat{\sigma}^2 = \frac{R_0^2}{107}. \tag{7.19}$$

The data has 109 degrees of freedom. Since two parameters $\widehat{\beta}_0, \widehat{\beta}_1$ are estimated, 2 degrees of freedom are lost and hence the effective sample size is 107. That is the reason for the denominator in (7.19). In the regression output produced by R (shown below), the square root of $\widehat{\sigma}^2$ is called the residual standard error, denoted as s_e.

Coefficient of Determination

How good is the fitted model in explaining the variation in the response variable, the adipose tissue area? The variation in the adipose tissue area can be represented by $\sum_{i=1}^{109} \left(AT_i - \overline{AT}\right)^2$. As we have seen above, the variation in adipose tissue area not explained by our model is given by R_0^2. Hence the part of variation in the adipose tissue area that is explained by our model is given by

$$\sum_{i=1}^{109} \left(AT_i - \overline{AT}\right)^2 - R_0^2. \tag{7.20}$$

Thus the proportion of the variation in the adipose tissue area that is explained by our model is

$$\frac{\sum_{i=1}^{109} \left(AT_i - \overline{AT}\right)^2 - R_0^2}{\sum_{i=1}^{109} \left(AT_i - \overline{AT}\right)^2} = 1 - \frac{R_0^2}{\sum_{i=1}^{109} \left(AT_i - \overline{AT}\right)^2} \tag{7.21}$$

This expression is called the coefficient of determination corresponding to the model and is denoted by R^2. Formally, R^2 is the ratio of the variation explained by the model to total variation in the response variable.

It is easy to see that $0 \leq R^2 \leq 1$ always. How do we interpret the extreme values for R^2?

If R^2 is 0, it means that there is no reduction in the variation in the response variable achieved by our model and thus this model is useless. (Caution: This however does not mean that the regressor is useless in explaining the variation in the response variable. It only means that the function, namely, the linear function in this case, is not useful. Some other function of the same regressor may be quite useful. See Exercise 7.1.)

On the other hand, if $R^2 = 1$, it means that $R_0^2 = 0$, which in turn means that each residual is 0. So the model fits perfectly to the data.

Let us recall that R^2 is *the proportion of variation in the response variable that is explained by the model.*

In the case of a *single regressor,* one can show that R^2 *is the square of the correlation coefficient between the response variable and the regressor* (see Exercise 7.2). This is the reason for saying that the correlation coefficient is a measure of the strength of a linear relationship between the response variable and the regressor (in the single regressor case).

However, the above two are the extreme cases. For almost all practical data sets, $0 < R^2 < 1$. Should we be elated when R^2 is large or should we be necessarily depressed when it is small? Fact is, R^2 is but just one measure of fit. We shall come back to this discussion later (see also Exercise 7.2).

Prediction for a New Observation

For a new individual whose waist circumference is available, say, $WC = x_0$ cm, how do we predict his abdominal adipose tissue? This is done by using the formula (7.16). Thus the predicted value of the adipose tissue for this person is

$$\widehat{AT} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \quad \text{sq.cm.} \tag{7.22}$$

with the standard error given by

$$s_1 = s_e \sqrt{1 + \frac{1}{107} + \frac{\left(x_0 - \overline{WC}\right)^2}{\sum_{i=1}^{109} \left(WC_i - \overline{WC}\right)^2}}. \tag{7.23}$$

The average value of the adipose tissue area for all the individuals with $WC = x_0$ cm is also estimated by the formula (7.22), with the standard error given by.

$$s_2 = s_e \sqrt{\frac{1}{107} + \frac{\left(x_0 - \overline{WC}\right)^2}{\sum_{i=1}^{109} \left(WC_i - \overline{WC}\right)^2}} \tag{7.24}$$

Notice the difference between (7.23) and (7.24). Clearly (7.23) is larger than (7.24). This is not surprising because the variance of an observation is larger than that of the average as seen in the Chap. 6, Statistical Methods—Basic Inferences. Why does one need the standard-error-formulae in (7.23) and (7.24)? As we see in Chap. 6, these are useful in obtaining the prediction and confidence intervals. Also note that the confidence interval is a statement of confidence about the true line—because we only have an estimate of the line. See (7.28) and (7.29) for details.

Testing of Hypotheses and Confidence Intervals

Consider the model (7.10). If $\beta_1 = 0$, then there is no linear relationship between AT and WC. Thus testing

$$H_0 : \beta_1 = 0 \quad against \quad H_1 : \beta_1 \neq 0 \tag{7.25}$$

is equivalent to testing for the usefulness of WC in predicting AT through a linear relationship. It can be shown that under the assumptions (7.13), $\widehat{\beta}_1$ as obtained in (7.15) has a normal distribution with mean β_1 and a suitable variance.

Thus, the test statistic to perform the test (7.25) is given by

$$\frac{\widehat{\beta}_1}{S.E.\left(\widehat{\beta}_1\right)} \tag{7.26}$$

Where $S.E.(.)$ stands for the standard error of (.).

As seen in Chap. 6, under the null hypothesis, (7.26) has a student's t distribution with 107 (109 minus 2) degrees of freedom. The corresponding p-value can be obtained as in Chap. 6. Testing $H_0 : \beta_1 = 0$ *against* $H_1 : \beta_1 \neq 0$ can be performed in a similar manner.

Again, as seen in Chap. 6, a 95% confidence interval for β_1 is given by

$$\left(\widehat{\beta}_1 - t_{107,0.025} S.E.\left(\widehat{\beta}_1\right), \widehat{\beta}_1 + t_{107,0.025} S.E.\left(\widehat{\beta}_1\right)\right) \tag{7.27}$$

where $t_{107,0.025}$ stands for the 97.5 percentile value of the student's t distribution with 107 degrees of freedom.

Similarly, a 95% confidence interval for the average value of the adipose tissue for all individuals having waist circumference equal to x_0 cm is given by

$$\left(A\widehat{T} - t_{107,0.025} s_2, A\widehat{T} + t_{107,0.025} s_2\right). \tag{7.28}$$

Also, a 95% *prediction interval* for the adipose tissue for an individual having waist circumference equal to x_0 cm is given by

$$\left(A\widehat{T} - t_{107,0.025} s_1, A\widehat{T} + t_{107,0.025} s_1\right). \tag{7.29}$$

From the expressions for $s_i, i = 1,2$ (as in 7.23 and 7.24), it is clear that the widths of the confidence and prediction intervals is the least when $x_0 = \overline{WC}$ and gets larger and larger as x_0 moves farther away from $\overline{WC}$. This is the reason why it is said that the prediction becomes unreliable if you try to predict the response variable value for a regressor value outside the range of the regressor values. The same goes for the estimation of the average response variable value.

Let us now consider the linear regression output in R for regressing AT on WC and interpret the same.

```
> model<-lm(AT ~ Waist, data=wc_at)
> summary(model)

Call:
lm(formula = AT ~ Waist, data = wc_at)

Residuals:
     Min       1Q   Median       3Q      Max
-107.288  -19.143   -2.939   16.376   90.342
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -215.9815    21.7963  -9.909   <2e-16 ***
Waist          3.4589     0.2347  14.740   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.06 on 107 degrees of freedom
Multiple R-squared:   0.67,     Adjusted R-squared:  0.667
F-statistic: 217.3 on 1 and 107 DF,  p-value: < 2.2e-16
```

Interpretation of the Regression Output

From the output, the following points emerge.

1. From the five-point summary (min, 1Q, median, etc.) and box plot of the residuals, it appears that the distribution of the residuals is skewed to the left. It also shows that there are four residuals which are too far from the center of the data. The normality of the residuals needs to be examined more closely.
2. The estimated regression equation is

$$\widehat{AT} = -215.9815 + 3.4589WC. \tag{7.30}$$

   From here we conclude that one cm increase in WC leads to an increase of 3.4589 sq.cm. in AT on average.
3. The t-value in each row is the ratio of the coefficient estimate and the standard error (see 7.26). The corresponding p-values are given in the next column of the table. The estimated coefficient estimates for both intercept and WC are highly significant. (The hypothesis $\beta_0 = 0$ against $\beta_0 \neq 0$ and also, the hypothesis $\beta_1 = 0$ against $\beta_1 \neq 0$ are both rejected since each of the p-values is smaller than 0.05.) This means that, based on this model, waist circumference does contribute to the variation in the adipose tissue area.
4. The estimate of the error-standard deviation, namely, the residual standard error is 33.06 cm.
5. The coefficient of determination R^2 is 0.67. This means that 67% of the variation in the adipose tissue area is explained by this model. When there is only one regressor in the model, this also means that the square of the correlation coefficient (0.8186) between the adipose tissue area and the waist circumference is 0.67. The sign of the correlation coefficient is the same as the sign of the regression coefficient estimate of WC which is positive. Hence, we conclude that AT, on average, increases with an increase in WC.
6. What is the interpretation of the F-statistic in the output? This is the ratio of the explained variation in the response variable to the residual or "still unexplained" variation based on the fitted model (after adjusting for the respective degrees of freedom). Intuitively, the larger this value, the better the fit because a large

part of the variation in the response variable is explained by the fitted model if the value is large. How does one judge how large is large? Statistically this statistic has an F distribution with the numerator and denominator degrees of freedom under the null hypothesis of the ineffectiveness of the model. In the single regressor case, the F-statistic is the square of the t-statistic. In this case, both the t-test for the significance of the regression coefficient estimate of WC and the F-statistic test for the same thing, namely, whether WC is useful for predicting AT (or, equivalently, in explaining the variation in AT) through the model under consideration.

7. Consider the adult males in the considered population with a waist circumference of 100 cm. We want to estimate the average abdominal adipose tissue area of these people. Using (7.30), the point estimate is $-215.9815 + 3.4589 \times 100 = 129.9085$ square centimeters.
8. Consider the same situation as in point 7 above. Suppose we want a 95% confidence interval for the average abdominal adipose tissue area of all individuals with waist circumference equal to 100 cm. Using the formula (7.28), we have the interval [122.5827, 137.2262]. Now consider a specific individual whose waist circumference is 100 cm. Using the formula (7.29), we have the 95% prediction interval for this individual's abdominal adipose tissue as [63.94943, 195.8595].
9. In point 7 above, if the waist circumference is taken as 50 cm, then using (7.30), the estimated average adipose tissue area turns out to be $-215.9815 + 3.4589 \times 50 = -42.9365$ square centimeters, which is absurd. Where is the problem? The model is constructed for the waist circumference in the range (63.5 cm, 119.90 cm). The formula (7.30) is applicable for estimating the average adipose tissue area when the waist circumference is in the range of waist circumference used in the estimation of the regression equation. If one goes much beyond this range, then the confidence intervals and the prediction intervals as constructed in the point above will be too large for the estimation or prediction to be useful.
10. Testing whether WC is useful in predicting the abdominal adipose tissue area through our model is equivalent to testing the null hypothesis, $\beta_1 = 0$. This can be done using (7.26) and this is already available in the output before Fig. 7.3. The corresponding p-value is 2×10^{-16}. This means that if we reject the hypothesis $\beta_1 = 0$, based on our data, then we reject wrongly only 2×10^{-16} proportion of times. Thus we can safely reject the null hypothesis and declare that WC is useful for predicting AT as per this model.

While we used this regression model of AT on Waist, this is not an appropriate model since the equal variance assumption is violated. For a suitable model for this problem, see Sect. 3.16.

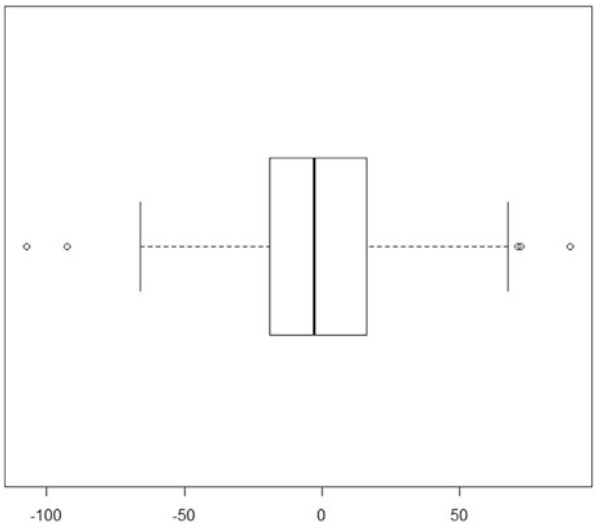


Fig. 7.3 Box plot of the residuals for the model AT vs WC

Table 7.2 Summary statistics of HP, MPG, and VOL

Variable	Min	1st quartile	Median	Mean	3rd quartile	Max.	Standard deviation
HP	49.0	84.0	100.0	117.5	140.0	322.0	57.1135
MPG	12.10	27.86	35.15	34.42	39.53	53.70	9.1315
VOL	50.00	89.00	101.00	98.77	113.00	160.00	22.3015

3.7 Multiple Regressors Case

Let us consider the example in Sect. 2.3. Here we have two regressors, namely, HP and VOL. As in the single regressor case, we can first look at the summary statistics and the box plots to understand the distribution of each variable (Table 7.2 and Fig. 7.4).

From the above summary statistics and plots, we notice the following:

(a) The distribution of MPG is slightly left skewed.
(b) The distribution of VOL is right skewed and there are two points which are far away from the center of the data (on this variable), one to the left and the other to the right.
(c) The distribution of HP is heavily right skewed.

Does point (a) above indicate a violation of the normality assumption (7.9)? Not necessarily, since the assumption (7.9) talks about the conditional distribution of MPG given VOL and HP whereas the box plot of MPG relates to the unconditional distribution of MPG. As we shall see later, this assumption is examined using residuals which are the representatives of the errors. Point (b) can be helpful when

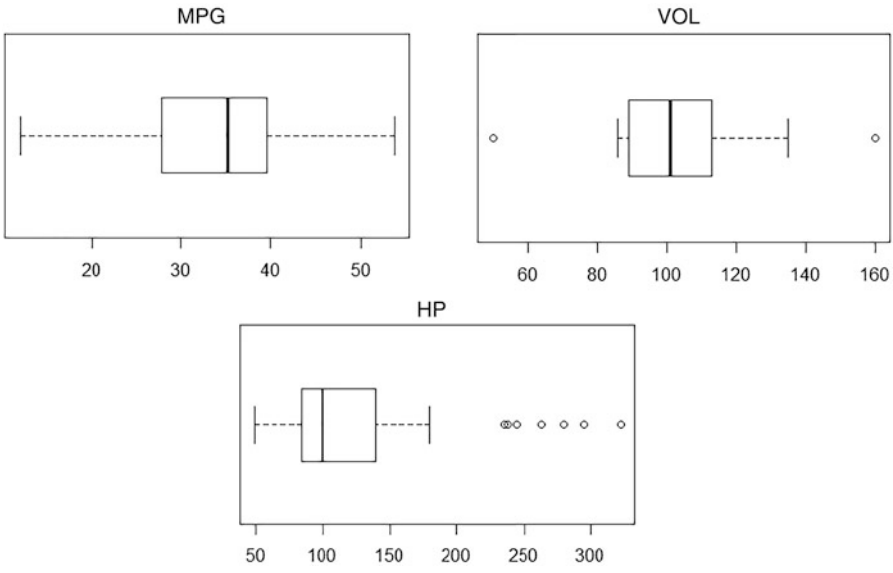


Fig. 7.4 Box plots of MPG, VOL, and HP

we identify, interpret and deal with unusual observations which will be discussed in the section. Point (c) will be helpful when we will look at the residual plots to identify suitable transformations which will be discussed in Sect. 3.10.

We can then look at the scatterplots for every pair of the variables: MPG, VOL, and HP. This can be put in the form of a matrix of scatterplots.

Scatterplots Matrix

The matrix of the scatterplots in Fig. 7.5, is called the scatterplot matrix of all the variables, namely, the response variable and the regressors. Unlike the scatterplot in the single regressor case, the scatterplot matrix in the multiple regressors case is of limited importance. In the multiple regressors case, we are interested in the influence of a regressor over and above that of other regressors. We shall elaborate upon this further as we go along. However, the scatterplots in the scatterplot matrix ignore the influence of the other regressors. For example, from the scatterplot of HP vs MPG (second row first column element in Fig. 7.2), it appears that MPG has a quadratic relationship with HP. But this ignores the impact of the other regressor on both MPG and HP. How do we take this into consideration? After accounting for these impacts, will the quadratic relationship still hold? We shall study these aspects in Sect. 3.10. The scatterplot matrix is useful in finding out whether there is almost perfect linear relationship between a pair of regressors. Why is this important? We shall study this in more detail in Sect. 3.13.

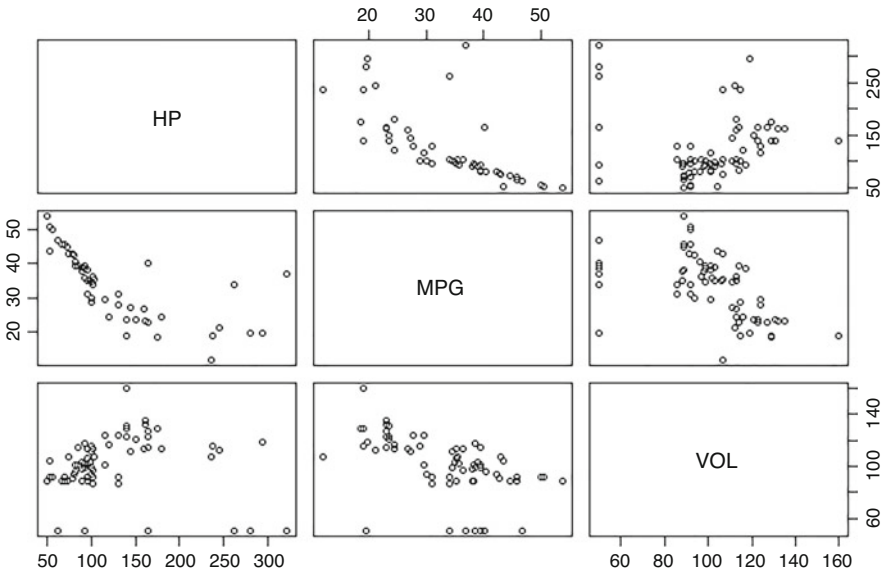


Fig. 7.5 Scatterplot matrix of variables

A linear regression model with two or more regressors is often referred to as a *multiple linear regression* model. Let us start with the following multiple linear regression model for the MPG problem.

$$\text{Model: } MPG = \beta_0 + \beta_1 HP + \beta_2 VOL + \varepsilon \tag{7.31}$$

Where does the error, ε, come from? We know that MPG is not completely determined by HP and VOL. For example, the age of the vehicle, the type of the road (smooth, bumpy, etc.) and so on impact the MPG. Moreover, as we have seen in the scatterplot: MPG vs HP, a quadratic term in HP may be warranted too. (We shall deal with this later.) All these are absorbed in ε.

$$\text{Data: } (MPG_i, HP_i, VOL_i), i = 1, \ldots, 81. \tag{7.32}$$

$$\text{Model adapted to data: } MPG_i = \beta_0 + \beta_1 HP_i + \beta_2 VOL_i + \varepsilon_i, i = 1, \ldots, 81 \tag{7.33}$$

Assumptions: $\varepsilon_i \mid HP_i, VOL_i, i = 1, \ldots, 81$ are independently and identically distributed as normal with mean 0 and variance σ^2 often written in brief as

$$\varepsilon_i \mid HP_i, VOL_i, i = 1, \ldots, 81 \text{ are } iid\ N\left(0, \sigma^2\right) \text{ variables.} \tag{7.34}$$

The model described by (7.33) and (7.34) is a special case of the model described by (7.7) and (7.9) where k = 2 and N = 81.

In the formula (7.32), β_1 and β_2 are the rates of change in MPG with respect to HP and VOL when the other regressor is kept fixed. Thus, these are partial rates of change. So strictly speaking, β_1 and β_2 should be called *partial regression coefficients*. When there is no confusion we shall refer to them as just regression coefficients.

Estimation of Parameters

As in the single regressor case, we estimate the parameters by *the method of least squares* (least sum of squares of the errors). Thus, we are led to the normal equations:

$$\left.\begin{array}{l} \sum_{i=1}^{81} MPG_i = 81\beta_0 + \beta_1 \sum_{i=1}^{81} HP_i + \beta_2 \sum_{i=1}^{81} VOL_i \\ \sum_{i=1}^{81} MPG_i HP_i = \beta_0 \sum_{i=1}^{81} HP_i + \beta_1 \sum_{i=1}^{81} HP_i^2 + \beta_2 \sum_{i=1}^{81} HP_i VOL_i \\ \sum_{i=1}^{81} MPG_i VOL_i = \beta_0 \sum_{i=1}^{81} VOL_i + \beta_1 \sum_{i=1}^{81} HP_i VOL_i + \beta_2 \sum_{i=1}^{81} VOL_i^2 \end{array}\right\} \tag{7.35}$$

The solution of the system of equations in (7.35) yields the estimates $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$ of the parameters β_0,β_1,β_2 respectively.

Thus, the estimated regression equation or the *fitted model* is

$$\widehat{MPG} = \widehat{\beta}_0 + \widehat{\beta}_1 HP + \widehat{\beta}_2 VOL \tag{7.36}$$

Along similar lines to the single regressor case, the *fitted value* and the *residual* corresponding to the i^{th} observation are, respectively,

$$\left.\begin{array}{l} \widehat{MPG}_i = \widehat{\beta}_0 + \widehat{\beta}_1 HP_i + \widehat{\beta}_2 VOL_i \\ e_i = MPG_i - \widehat{MPG}_i \end{array}\right\}. \tag{7.37}$$

We recall that the fitted value and the residual for the i^{th} observation are the explained and the unexplained parts of the observation based on the fitted model.

Residual Sum of Squares

As before, we use their sample representatives e_i to check the assumptions (7.9) on the errors.

The sum of squares of residuals, $R_0^2 = \sum_{i=1}^{81} e_i^2$ is the part of the variation in MPG that is not explained by the fitted model (7.36) obtained by the method of least squares.

The estimator of σ^2 is obtained as

$$\widehat{\sigma}^2 = \frac{R_0^2}{78}. \tag{7.38}$$

The data has 81 degrees of freedom. Since three parameters $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$ are estimated, three degrees of freedom are lost and hence the effective sample size is $81 - 3 = 78$. That is the reason for the denominator in (7.38). As mentioned earlier, the square root of $\widehat{\sigma}^2$ is called the residual standard error in the R package.

Interpretation of the Regression Coefficient Estimates and Tests of Significance

From (7.37), we see that $\widehat{\beta}_1$ is the estimated change in MPG per unit increase in HP when VOL is kept constant. (It is easy to see that $\widehat{\beta}_1$ is the partial derivative, $\frac{\partial \widehat{MPG}}{\partial HP}$.) Notice the change in the interpretation of $\widehat{\beta}_1$ from that of the corresponding coefficient estimate in the single regressor case. The significance of $\widehat{\beta}_1$ is tested in the same way as in the single regressor case, that is, using the statistic as in (7.26). The degrees of freedom for the t-statistic are the same as those of the residual sum of squares, namely, 78 (the total number of observations minus the number of β parameters estimated).

Coefficient of Multiple Determination, R^2

As in the single regressor case, we ask the question: How good is the fitted model in explaining the variation in the response variable, MPG? The variation in MPG can be represented by $\sum_{i=1}^{81}\left(MPG_i - \overline{MPG}\right)^2$. As we saw, the variation in MPG not explained by our fitted model is given by R_0^2. Hence the part of variation in MPG that is explained by our model is given by

$$\sum_{i=1}^{81}\left(MPG_i - \overline{MPG}\right)^2 - R_0^2. \tag{7.39}$$

Thus the proportion of the variation in MPG that is explained by our fitted model is

$$\frac{\sum_{i=1}^{81}\left(MPG_i - \overline{MPG}\right)^2 - R_0^2}{\sum_{i=1}^{81}\left(MPG_i - \overline{MPG}\right)^2} = 1 - \frac{R_0^2}{\sum_{i=1}^{81}\left(MPG_i - \overline{MPG}\right)^2} \tag{7.40}$$

This expression, similar to that in (7.21) is called the *coefficient of multiple determination* corresponding to the fitted model and is also denoted by R^2.

It is easy to see that $0 \leq R^2 \leq 1$ always. The interpretation of the extreme values for R^2 is the same as in the single regressor case.

Adjusted R^2

It can be shown that R^2 almost always increases with more regressors. (It never decreases when more regressors are introduced.) So R^2 may not be a very good criterion to judge whether a new regressor should be included. So it is meaningful to look for criteria which impose a penalty for unduly bringing in a new regressor into the model. One such criterion is Adjusted R^2 defined below. (When we deal with subset selection, we shall introduce more criteria for choosing a good subset of regressors.) We know that both R_0^2 and $\sum_{i=1}^{81}\left(MPG_i - \overline{MPG}\right)^2$ are representatives of the error variance σ^2. But the degrees of freedom for the former is $81 - 3 = 78$ and for the latter is $81 - 1 = 80$. When we are comparing both of them as in (7.40),

some feel that we should compare the measures per unit degree of freedom as then they will be true representatives of the error variance. Accordingly, adjusted R^2 is defined as

$$\text{Adj } R^2 = 1 - \frac{R_0^2/(81-3)}{\sum_{i=1}^{81}\left(MPG_i - \overline{MPG}\right)^2/(81-1)}. \tag{7.41}$$

The adjusted R^2 can be written as $1 - \frac{n-1}{n-K}\left(1-R^2\right)$, where n is the number of observations and K is the number of parameters of our model that are being estimated. From this it follows, that unlike R^2, adjusted R^2may decrease with the introduction of a new regressor. However, adjusted R^2 may become negative and does not have the same intuitive interpretation as that of R^2.

As we can see, adjusted R^2 is always smaller than the value of R^2. A practical thumb rule is to examine whether R^2 and adjusted R^2 are quite far apart. One naïve way to judge this is to see if the relative change, $\frac{\text{R}^2\text{-adj R}^2.}{\text{R}^2}$, is more than 10%. If it is not, go ahead and interpret R^2. However, if it is, then it is an indication that there is some issue with the model—either there is an unnecessary regressor in the model or there is some unusual observation. We shall talk about the unusual observations in Sect. 3.12.

Let us now consider Example 2.3, the gasoline consumption problem. We give the R-output of the linear regression of MPG on HP and VOL in Table 7.3.

Table 7.3 Regression output

```
> model1<-lm(MPG ~ HP + VOL, data=Cars)
> summary(model1)

Call:
lm(formula = MPG ~ HP + VOL, data = Cars)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3128 -3.3714 -0.1482  2.8260 15.4828

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.586385   2.505708  26.574  < 2e-16 ***
HP          -0.110029   0.009067 -12.135  < 2e-16 ***
VOL         -0.194798   0.023220  -8.389 1.65e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.618 on 78 degrees of freedom
Multiple R-squared:  0.7507,   Adjusted R-squared:  0.7443
F-statistic: 117.4 on 2 and 78 DF,  p-value: < 2.2e-16
```

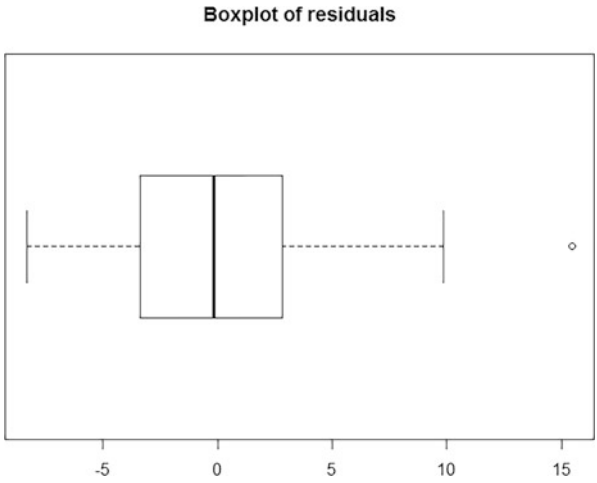


Fig. 7.6 The box plot of the residuals for the model MPG vs VOL and HP

Interpretation of the Regression Output

From the output and the plot above, we notice the following:

1. From the five-point summary of the residuals and the box-plot in Fig. 7.6, we observe that the distribution of the residuals is skewed to the right. Also, there appears to be one unusually large residual which is marked by a small circle towards the right in the box plot. This may have an implication towards the deviation from normality of the residuals. We shall examine this more closely using the QQ plot of suitable residuals.
2. The estimated regression equation is:

$$\widehat{MPG} = 66.586385 - 0.110029HP - 0.194798VOL. \tag{7.42}$$

3. Do the signs of the regression coefficient estimates conform to the intuition? Higher horse power cars consume more petrol and hence give lower MPG. The negative sign of the regression coefficient conforms to this intuition. Similarly large cars consume more petrol justifying the negative coefficient estimate of VOL.
4. From (7.42) we infer that a 10 cubic feet increase in the volume of the vehicle, with no change in horse power, will lead to a decrease of approximately 2 miles (to be precise, $10 \times 0.194798 = 1.94798$ miles) per gallon. Notice the difference in the interpretation of the (partial) regression coefficient estimate of VOL from the interpretation of a regression coefficient estimate in a single regressor case. In the present case it is the rate of change in the response variable with respect to the regressor under consideration, keeping the other regressors fixed. From (7.42), a unit increase in HP, keeping the volume unchanged, will lead to a reduction in the mileage, that is, a reduction in MPG by 0.110029. This is based on the data on the sample collected, if another sample is used, then the coefficient estimate

of HP is unlikely to be exactly the same as 0.110029. So, can we give some realistic approximate bounds for the coefficient β_1 of HP in the population? In other words, we are seeking an interval estimate of the coefficient. A 95% confidence interval for the coefficient of HP is given by

$$\widehat{\beta_1} \pm t_{78,0.025} S.E.\left(\widehat{\beta_1}\right).$$

The computed 95% confidence interval for β_1 is $(-0.1280797, -0.09197832)$. Thus in the population of the vehicles, a unit increase in the horse power with no change in volume can lead to as high a decrease in the MPG as 0.128 or as low as 0.092. Such a statement can be made with 95% confidence.

5. The tests for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ and $H_2 : \beta_2 = 0$ against $H_3 : \beta_2 \neq 0$ can be performed in the same way as in the single regressor case (see point 3 in the interpretation of the output in Sect. 3.6). What do these tests mean and why are they important? The estimated coefficients of HP and VOL are different from 0. But these are estimated from the sample under consideration. If we conclude that the coefficient of HP is different from 0, what are the chances that we are wrong? Since the p-value for rejecting H_0 against H_1 is less than 2×10^{-16}, the probability that we are wrong is less than 2×10^{-16} which is very small. So we can conclude safely that the coefficient of HP in the population is different from 0. Likewise, we can conclude that the coefficient of VOL is also different from 0. If the (partial) regression coefficient of HP were 0, it would mean that HP cannot at all explain the variation in MPG over and above whatever is explained by VOL. Thus in the presence of VOL, HP is not useful in explaining the variation in MPG. But that is not the case in this instance. From our results, we find that both HP (respectively VOL) is useful in explaining the variation in MPG in the presence of VOL (respectively HP).
6. The usual point estimate of the average value of MPG for all vehicles with HP = 150 and VOL = 100 cubic feet can be obtained using (7.12) and is given by

$$\widehat{MPG} = 66.586385 - 0.110029 \times 150 - 0.194798 \times 100 = 30.602235.$$

7. A 95% confidence interval for the average value of MPG for all vehicles with HP = 150 and VOL = 100 cubic feet can be obtained as follows.

   Let Z be the matrix of order 81×3 where each element in the first column is 1, the second and third columns, respectively, are data on HP and VOL on vehicles 1 to 81 in that order. Let $u^t = \begin{pmatrix} 1 & 150 & 100 \end{pmatrix}$. Then the required confidence interval is given by the formula

$$\left(u^t\widehat{\beta} - \sqrt{u^t(Z^tZ)^{-1}u\widehat{\sigma}^2}\, t_{78,0.025},\, u^t\widehat{\beta} + \sqrt{u^t(Z^tZ)^{-1}u\widehat{\sigma}^2}\, t_{78,0.025}\right)$$

where $\widehat{\sigma}^2$, the estimate of the error variance, is the square of the residual standard error in the output. The computed interval is (29.42484, 31.77965).

A 95% prediction interval for MPG of a vehicle with HP = 150 and VOL = 100 cubic feet can be obtained as follows. Let Z, u, and $\widehat{\sigma}^2$ be as specified in point 8 above. Then the required prediction interval is given by the formula:

$$\left(u^t\widehat{\beta}-\sqrt{\left(1+u^t(Z^tZ)^{-1}u\right)\widehat{\sigma}^2}t_{78,0.025},\, u^t\widehat{\beta}+\sqrt{\left(1+u^t(Z^tZ)^{-1}u\right)\widehat{\sigma}^2}t_{78,0.025}\right).$$

The computed interval is (21.33388, 39.87062). Notice the difference in the confidence and prediction intervals. The prediction interval is always larger than the corresponding confidence interval. Do you know why? The variation in the average is always smaller than the individual variation.

8. The coefficient of multiple determination, R^2 is 75.07%. This means that HP and VOL together explain 75.07% of the variation in MPG through the current model. The adjusted R square is 74.43%. Since the relative change $\frac{\text{R}^2\text{-adj R}^2.}{\text{R}^2}$ is much smaller than 10%, we go on to interpret the R^2 as described above.
9. What does the F-test do here? It tests whether the regressors used, namely, HP and VOL together, have any explaining capacity regarding the variance in MPG, through the model under consideration. If the F-test turns out to be insignificant, then this model is not worth considering. However, it is wrong to conclude that the regressors themselves are not useful for predicting MPG. Perhaps some other model using the same regressors may yield a different result.

3.8 Why Probe Further?

We have so far considered the cases of single regressor and multiple regressors in Linear Regression and estimated the model which is linear in parameters and also in regressors. We interpreted the outputs and obtained the relevant confidence and prediction intervals. We also performed some basic tests of importance. Are we done? First, let us consider the data sets of Anscombe (1973) as given in Table 5.1 and Fig. 5.3 of the data visualization chapter (Chap. 5). Look at Table 5.1. There are four data sets, each having data on two variables. The plan is to regress y_i on x_i, $i = 1, \ldots, 4$. From the summary statistics, we notice the means and standard deviations of the x's are the same across the four data sets and the same is true for the y's. Further the correlation coefficient in each of the four data sets are the same. Based on the formulae (7.15) adapted to these data sets, the estimated regression lines are the same. Moreover, the correlation coefficient is 0.82 which is substantial. So it appears that linear regression is equally meaningful in all the four cases and gives the same regression line.

Now let us look at the plots in the Fig. 5.3. For the first data set, a linear regression seems reasonable. From the scatterplot of the second data set, a parabola is more appropriate. For the third data set, barring the third observation (13, 12.74), the remaining data points lie almost perfectly on a straight line, different from the line in the plot. In the fourth data set, the x values are all equal to 8, except for the eighth observation where the x value is 19. The slope is influenced by this observation. Otherwise we would never use these data (in the fourth data set) to predict y based on x.

This reiterates the observation made earlier: One should examine suitable plots and other diagnostics to be discussed in the following sections before being satisfied with a regression. In the single regressor case, the scatterplot would reveal a lot of useful information as seen above. However, if there are several regressors, scatterplots alone will not be sufficient since we want to assess the influence of a regressor on the response variable after controlling for the other regressors. Scatterplots ignore the information on the other regressors. As we noticed earlier, the residuals are the sample representatives of the errors. This leads to the examination of some suitable residual plots to check the validity of the assumptions in Eq. (7.9). In the next section, we shall develop some basic building blocks for constructing the diagnostics.

3.9 Leverage Values, Residuals, and Outliers

The Hat Matrix: First, we digress a little bit to discuss a useful matrix. Consider the linear regression model (7.8) with the assumptions (7.9). There is a matrix H, called the *hat matrix* which transforms the vector of response values Y to the vector of the fitted values $\widehat{Y}$. In other words, $\widehat{Y} = HY$. The hat matrix is given by $\mathrm{Z}(\mathrm{Z}^{\mathrm{t}}\mathrm{Z})^{-1}\mathrm{Z}^{\mathrm{t}}$. It is called the hat matrix because when applied to Y it gives the estimated (or hat) value of Y. The hat matrix has some good properties, some of which are listed below

(a) $Var\left(\widehat{Y}\right) = \sigma^2 H$.
(b) $Var(residuals) = \sigma^2(I - H)$, where I is the identity matrix.

Interpretation of Leverage Values and Residuals Using the Hat Matrix

Diagonal elements h_{ii} of the hat matrix have a good interpretation. If h_{ii} is large, then the regressor part of data for the i^{th} observation is far from the center of the regressor data. If there are N observations and k regressors, then h_{ii} is considered to be large if it is $\frac{2(k+1)}{N}$ or higher.

Let us now look at the residuals. If the residual corresponding to the i^{th} residual is large, then the fit of the i^{th} observation is poor. To examine whether a residual is large, we look at a standardized version of the residual. It can be shown that the mean of each residual is 0 and the variance of the i^{th} residual is given by $Var(e_i) = (1 - h_{ii})\sigma^2$. We recall that σ^2 is the variance of the error of an observation (of course conditional on the regressors). The estimate of σ^2 obtained from the

model after dropping the i^{th} observation, denoted by $\widehat{\sigma}^2_{(i)}$ is preferred (since large i^{th} residual also has a large contribution to the residual sum of squares and hence to the estimate of σ^2).

An observation is called an *Outlier* if its fit in the estimated model is poor, or, equivalently, if its residual is large.

The statistic that is used to check whether the i^{th} observation is an outlier is

$$r_i = \frac{e_i - 0}{\sqrt{(1 - h_{ii})\,\widehat{\sigma}^2_{(i)}}}, \tag{7.43}$$

often referred to as the i^{th} studentized residual, which has a t distribution with $N - k - 1$ degrees of freedom under the null hypothesis that the i^{th} observation is not an outlier.

When one says that an observation is an outlier, it is in the context of the estimated model under consideration. The same observation may be an outlier with respect to one model and may not be an outlier in a different model (see Exercise 7.3).

As we shall see, the residuals and the leverage values form the basic building blocks for the deletion diagnostics, to be discussed in Sect. 3.12.

3.10 Residual Plots

We shall now proceed to check whether the fitted model is adequate. This involves the checking of the assumptions (7.9). If the fitted model is appropriate then the residuals are uncorrelated with the fitted values and also the regressors. We examine these by looking at the residual plots:

(a) Fitted values vs residuals
(b) Regressors vs residuals

In each case the residuals are plotted on the Y-axis. If the model is appropriate, then each of the above plots should yield a random scatter. The deviation from the random scatter can be tested, and an R package command for the above plots gives the test statistics along with the p-values.

Let us look at the residual plots and test statistics for the fitted model for the gasoline consumption problem given in Sect. 3.7.

How do we interpret Fig. 7.7 and Table 7.4? Do they also suggest a suitable corrective action in case one such is warranted?

We need to interpret the figure and the table together.

We notice the following.

1. The plot HP vs residual does not appear to be a random scatter. Table 7.4 also confirms that the p-value corresponding to HP is very small. Furthermore, if we

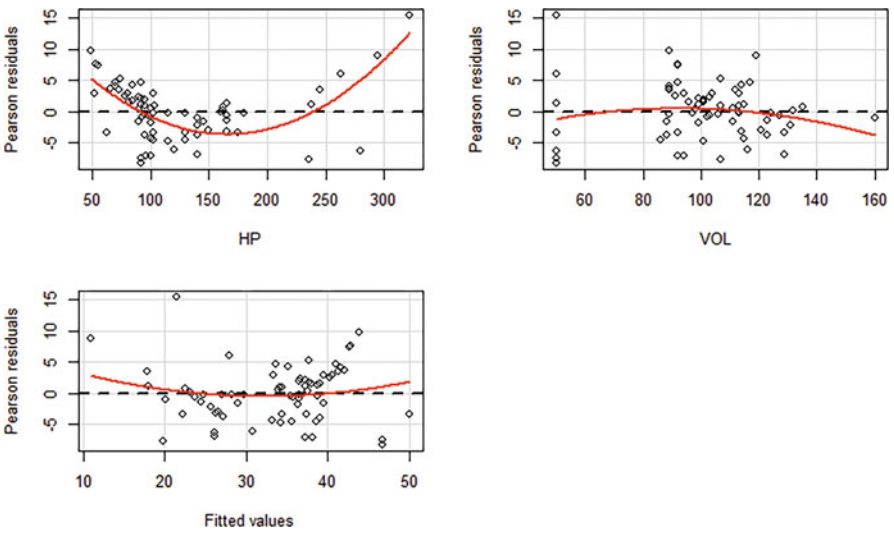


Fig. 7.7 The residual plots for gasoline consumption model in Sect. 3.7

Table 7.4 The tests for deviation from random scatter

```
> residualPlots(model1)

           Test stat Pr(>|t|)
HP            10.042    0.000
VOL           -1.575    0.119
Tukey test     1.212    0.226
```

look at the plot more closely we can see a parabolic pattern. Thus the present model is not adequate. We can try to additionally introduce the square term for HP to take care of this parabolic nature in the plot.

2. The plot VOL vs residual is not quite conclusive whether it is really a random scatter. Read this in conjunction with Table 7.4. The p-value corresponding to VOL is not small. So we conclude that there is not a significant deviation from a random scatter.
3. Same is the case with the fitted values vs residual. For fitted values vs residual, the corresponding test is the Tukey test.

A note of caution is in order. The R package command for the residual plots leads to plots where a parabolic curve is drawn to notify a deviation from random scatter. First, the deviation from random scatter must be confirmed from the table of tests for deviation from random scatter by looking at the p-value. Second, even if the p-value is small, it does not automatically mean a square term is warranted. Your judgment of the plot is important to decide whether a square term is warranted or something else is to be done. One may wonder why it is important to have the plots if, anyway, we need to get the confirmation from the table of tests for random scatter. The test only tells us whether there is a deviation from random scatter but it does not guide us to what transformation is appropriate in case of deviation. See,

for example, the residual plot of HP vs Residual in Fig. 7.7. Here we clearly see a parabolic relationship indicating that we need to bring in HP square. We shall see more examples in the following sections.

Let us add the new regressor which is the square of HP and look at the regression output and the residual plots.

The following points emerge:

1. From Table 7.5, we notice that HP_Sq is highly significant (p-value: 1.2×10^{-15}) and is positive.
2. R square and Adj. R square are pretty close. So we can interpret R square. The present model explains 89.2% of the variation in MPG. (Recall that the model in Sect. 3.7 explained only 75% of the variation in MPG.)
3. Figure 7.8 and Table 7.6 indicate that the residuals are uncorrelated with the fitted values and regressors.
4. Based on Table 7.5, how do we assess the impact of HP on MPG? The partial derivative of $\widehat{MPG}$ with respect to HP is $-0.4117 + 0.001808$ HP. Unlike in the model in Sect. 3.7 (see point 4), in the present model, the impact of unit increase in HP on the estimated MPG, keeping the VOL constant, depends on the level of HP. At the median value of HP (which is equal to 100—see Table 7.2), one unit increase in HP will lead to a reduction in $\widehat{MPG}$ by 0.2309 when VOL is kept constant. From the partial derivative, it is clear that as long as HP is smaller than 227.710177, one unit increase in HP will lead to a reduction in $\widehat{MPG}$, keeping VOL constant. If HP is greater than this threshold, then based on this model, $\widehat{MPG}$ will increase (happily!) with increasing HP when the VOL is held constant.

The question is: Are we done? No, we still need to check a few other assumptions, like normality of the errors. Are there some observations which are driving the results? Are there some important variables omitted?

Suppose we have performed a linear regression with some regressors. If the plot of fitted values vs residuals shows a linear trend, then it is an indication that there is an omitted regressor. However, it does not give any clue as to what this regressor is. This has to come from domain knowledge. It may also happen that, after the regression is done, we found another variable which, we suspect, has an influence on the response variable. In the next section we study the issue of bringing in a new regressor.

3.11 Added Variable Plot and Partial Correlation

Let us consider again the gasoline consumption problem. Suppose we have run a regression of MPG on VOL. We feel that HP also has an explaining power of the variation in MPG. Should we bring in HP? A part of MPG is already explained by VOL. So the unexplained part of MPG is the residual e (unexplained part of MPG) after regressing MPG on VOL. There is a residual value corresponding to

Table 7.5 Regression output of MPG vs HP, HP_SQ (Square of HP), and VOL

```
> model2<-lm(MPG ~ HP + VOL + HP_sq, data=Cars)
> summary(model2)

Call:
lm(formula = MPG ~ HP + VOL + HP_sq, data = Cars)

Residuals:
   Min     1Q Median     3Q    Max
-8.288 -2.037  0.561  1.786 11.008

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.744e+01  1.981e+00  39.100  < 2e-16 ***
HP          -4.117e-01  3.063e-02 -13.438  < 2e-16 ***
VOL         -1.018e-01  1.795e-02  -5.668  2.4e-07 ***
HP_sq        9.041e-04  9.004e-05  10.042  1.2e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.058 on 77 degrees of freedom
Multiple R-squared:  0.892,    Adjusted R-squared:  0.8878
F-statistic: 212.1 on 3 and 77 DF,  p-value: < 2.2e-16
```

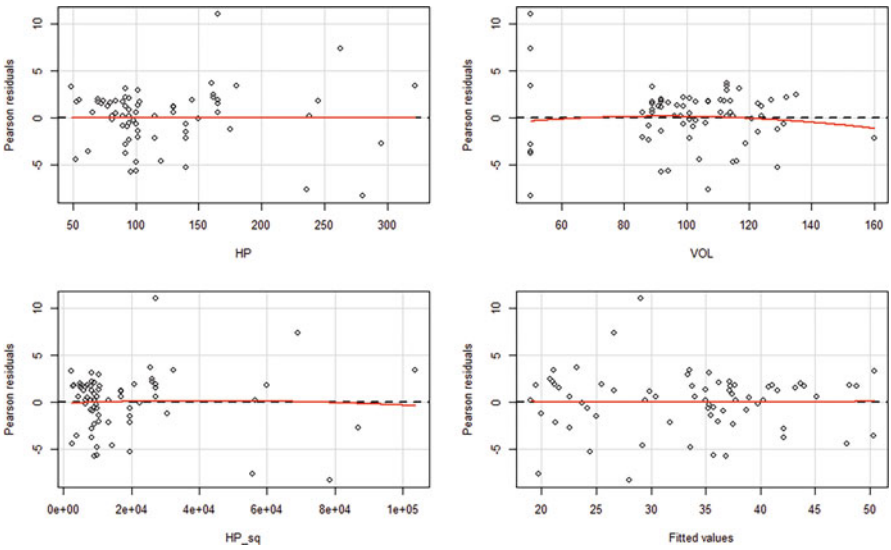


Fig. 7.8 Residual plots corresponding to the fitted model in Table 7.5

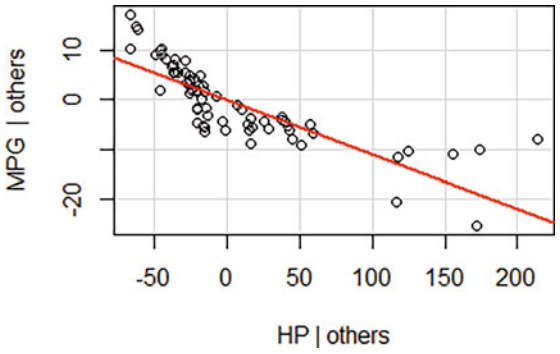
each observation. Let us call these residual values e_i, $i = 1, \ldots, 81$. So we will bring in HP if it has an explaining power of this residual. At the same time, a part of the explaining power of HP may already have been contained in VOL. Therefore, the

Table 7.6 The tests for deviation from uncorrelatedness for the fitted model in Table 7.5

```
> residualPlots(model2)

           Test stat Pr(>|t|)
HP             0.371    0.712
VOL           -0.703    0.484
HP_sq         -0.528    0.599
Tukey test     0.089    0.929
```

Fig. 7.9 Added variable plot



part of HP that is useful for us is the residual f after regressing HP on VOL. Let us call these residuals f_i,$i = 1, \ldots, 81$. It now boils down to the single regressor case of regressing e on f. It is natural to look at the scatterplot (f_i, e_i),$i = 1, \ldots, 81$. This scatterplot is called the *Added Variable Plot*. The correlation coefficient is called the *partial correlation coefficient* between MPG and HP fixing (or equivalently, eliminating the effect of) VOL. Let us look at this added variable plot (Fig. 7.9).

The following points emerge from the above added variable plot.

(a) As the residual of HP increases, the residual of MPG decreases, by and large. Thus it is useful to bring in HP into the regression in addition to VOL.

(b) We can notice a parabolic trend in the plot, suggesting that HP be brought in using a quadratic function. Thus, it reinforces our earlier decision (see Sect. 3.10.) to use HP and HP_SQ. Barring exceptions, in general, the added variable plot suggests a suitable functional form in which the new regressor can be added.

(c) The partial correlation coefficient is the strength of a linear relationship between the two residuals under consideration. For the gasoline consumption example, the partial correlation coefficient between MPG and HP keeping VOL fixed is -0.808. The correlation coefficient between MPG and HP is -0.72. (As we know the negative sign indicates that as HP increases, MPG decreases.) Is it surprising?

It is a common misconception that the correlation coefficient between two variables is always larger than or equal to the partial correlation coefficient between these variables after taking away the effect of other variables. Notice that the correlation coefficient between MPG and HP represents the strength of

the linear relationship between these variables ignoring the effect of VOL. The partial correlation coefficient between MPG and HP after taking away the effect of VOL is the simple correlation coefficient between the residuals e and f which are quite different from MPG and HP, respectively. *So a partial correlation coefficient can be larger than or equal to or less than the corresponding correlation coefficient.*

We give an interesting formula for computing the partial correlation coefficient using just the basic regression output. Let T denote the t-statistic value for a regressor. Let there be N observations and $k(\geq 2)$ regressors plus one intercept. Then the partial correlation coefficient between this regressor and the response variable (Greene 2012, p. 77) is given by

$$\frac{T^2}{T^2 + (N - k - 1)}. \tag{7.44}$$

Recall (see 7.38) that $N - k - 1$ is the degree of freedom for the residual sum of squares.

3.12 Deletion Diagnostics

Look at the third data set of Anscombe (1973) described in Sect. 3.8 and its scatterplot (Fig. 5.3) in the Data Visualization chapter (Chap. 5). But for the third observation (13, 12.74), the other observations fall almost perfectly on a straight line. Just this observation is influencing the slope and the correlation coefficient. In this section, we shall explain what we mean by an influential observation, give methods to identify such observations, and finally discuss what one can do with an influential observation.

We list below some of the quantities of interest in the estimation of a linear regression model.

1. Regression coefficient estimates
2. Fit of an observation
3. Standard errors of the regression coefficient estimates
4. The error variance
5. Coefficient of multiple determination

In a linear regression model (7.8) with the assumptions as specified in (7.9), no single observation has any special status. If the presence or absence of a particular observation can make a large (to be specified) difference to some or all of the quantities above, we call such an observation an *influential observation*.

Let us describe some notation before we give the diagnostic measures. Consider the model specified by (7.8) and (7.9) and its estimation in Sect. 3.7. Recall the definitions of the fitted values, $\widehat{y}_i$, the residuals, e_i, the residual sum of squares, R_0^2,

and the coefficient of multiple determination, R^2, given in Sect. 3.7. Let $\widehat{\beta}$ denote the vector of the estimated regression coefficients.

In order to assess the impact of an observation on the quantities (1)–(5) mentioned above, we set aside an observation, say the i^{th} observation and estimate the model with the remaining observations. We denote the vector of estimated regression coefficients, fitted value of the j^{th} observation, residual of the j^{th} observation, residual sum of squares and the coefficient of multiple determination after dropping the i^{th} observation by $\widehat{\beta}_{(i)}, \widehat{y}_{j(i)}, e_{j(i)}, R^2_{0(i)}, R^2_{(i)}$ respectively.

We give below a few diagnostic measures that are commonly used to detect influential observations.

(a) *Cook's distance*: This is an overall measure of scaled difference in the fit of the observations due to dropping an observation. This is also a scaled measure of the difference between the vectors of regression coefficient estimates before and after dropping an observation. More specifically, Cook's distance after dropping the i^{th} observation, denoted by $Cookd_i$, is proportional to $\sum_{j=1}^{N}\left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2$ where N is the number of observations. ($Cookd_i$ is actually a squared distance.) The i^{th} observation is said to be influential if $Cookd_i$ is large. If $Cookd_i$ is larger than a cutoff value (usually 80^{th} or 90^{th} percentile value of F distribution with parameters k and $N - k - 1$ where N is the number of observations and k is the number of regressors), then the i^{th} observation is considered to be influential. In practice, a graph is drawn with an observation number in the X-axis and Cook's distance in the Y-axis, called the index plot of Cook's distance, and a few observations with conspicuously large Cook's distance values are treated as influential observations.

(b) $DFFITS_i$: This is a scaled absolute difference in the fits of the i^{th} observation before and after the deletion of the i^{th} observation. More specifically, $DFFITS_i$ is proportional to $\mid \widehat{y}_i - \widehat{y}_{i(i)} \mid$. Observations with *DFFITS* larger than $2\sqrt{\frac{k+1}{N}}$ are flagged as influential observations.

(c) $COVRATIO_i$: $COVRATIO_i$ measures the change in the overall variability of the regression coefficient estimates due to the deletion of the i^{th} observation. More specifically it is the ratio of the determinants of the covariance matrices of the regression coefficient estimates after and before dropping the i^{th} observation. If $\mid COVRATIO_i - 1 \mid > \frac{3(k+1)}{N}$, then the i^{th} observation is flagged as an influential observation in connection with the standard errors of the estimates. It is instructive to also look at the index plot of *COVRATIO*.

(d) The scaled residual sum of squares estimates the error variance as we have seen in (7.38). The difference in the residual sum of squares R^2_0 and $R^2_{0(i)}$ before and after deletion of the i^{th} observation, respectively, is given by

$$R^2_0 - R^2_{0(i)} = \frac{e_i^2}{1 - h_{ii}}.$$

Thus the i^{th} observation is flagged as influential in connection with error variance if it is an outlier (see 7.43).

Two points are worth noting:

(a) If an observation is found to be influential, it does not automatically suggest "off with the head." The diagnostics above are only markers suggesting that an influential observation has to be carefully examined to find out whether there is an explanation from the domain knowledge and the data collection process why it looks different from the rest of the data. Any deletion should be contemplated only after there is a satisfactory explanation for dropping, from the domain knowledge.

(b) The diagnostics are based on the model developed. If the model under consideration is found to be inappropriate otherwise, then these diagnostics are not applicable.

We shall illustrate the use of Cook's distance using the following example on cigarette consumption. The data set "CigaretteConsumption.csv" is available on the book's website.

Example 7.1. A national insurance organization in USA wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. The variables chosen for the study are given in Fig. 7.10.

Variable	Definition
Age	Median age of a person living in a state
HS	% of people over 25 years of age in a state who completed high school
Income	Per capita personal income in a state (in dollars)
Black	% of blacks living in a state
Female	% of females living in a state
Price	Weighted average price (in cents) of a pack of cigarettes in a state
Sales	Number of packs of cigarettes sold in a state on a per capita basis

The R output of the regression of Sales on the other variables is in Table 7.7.

The index plots of Cook's distance and studentized residuals are given in Fig. 7.10.

From the Cook's distance plot, observations 9, 29, and 30 appear to be influential. Observations 29 and 30 are also outliers. (These also are influential with respect to error variance.) On scrutiny, it turns out that observations 9, 29, and 30 correspond to Washington DC, Nevada, and New Hampshire, respectively. Washington DC is the capital city and has a vast floating population due to tourism and otherwise. Nevada is different from a standard state because of Las Vegas. New Hampshire does not impose sales tax. It does not impose income tax at state level. Thus these three states behave differently from other states with respect to cigarette consumption. So it is meaningful to consider regression after dropping these observations.

The corresponding output is provided in Table 7.8.

Table 7.7 Regression results for cigarette consumption data

```
> model3<-lm(Sales ~ ., data=CigaretteConsumption[,-1])
> summary(model3)

Call:
lm(formula = Sales ~ ., data = CigaretteConsumption[, -1])

Residuals:
    Min      1Q  Median      3Q     Max
-48.398 -12.388  -5.367   6.270 133.213

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.34485  245.60719   0.421  0.67597
Age           4.52045    3.21977   1.404  0.16735
HS           -0.06159    0.81468  -0.076  0.94008
Income        0.01895    0.01022   1.855  0.07036 .
Black         0.35754    0.48722   0.734  0.46695
Female       -1.05286    5.56101  -0.189  0.85071
Price        -3.25492    1.03141  -3.156  0.00289 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.17 on 44 degrees of freedom
Multiple R-squared:  0.3208,   Adjusted R-squared:  0.2282
F-statistic: 3.464 on 6 and 44 DF,  p-value: 0.006857
```

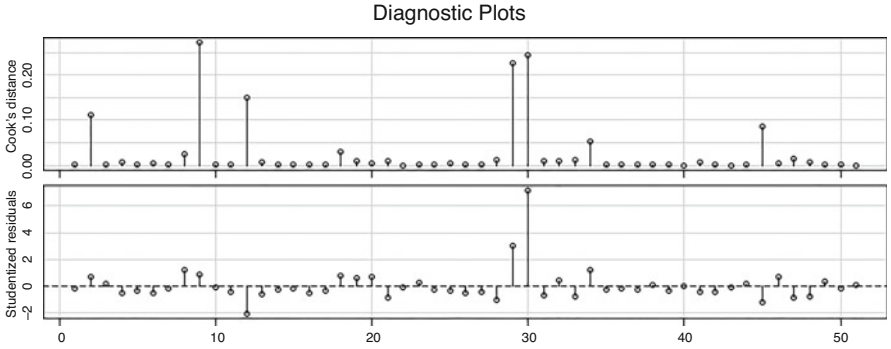


Fig. 7.10 Index plots for cigarette consumption model in Table 7.7

Notice the appreciable changes in the regression coefficient estimates and standard errors. The coefficients of HS and Price changed from −0.062 and −3.255 to −1.172 and −2.782, respectively. While the income coefficient estimate has not changed very much (from 0.019 to 0.021), the standard error got almost halved from 0.010 to 0.005, thereby Income became highly significant from being insignificant at 5% level. There are also changes in other coefficient estimates (including changes in sign), but we are not emphasizing them since they are significant in both the

Table 7.8 Regression results for cigarette consumption data after dropping observations 9, 29, and 30

```
> model4<- lm (Sales ~ ., data = CigaretteConsumption[-c(9,29,30),-1])
> summary(model4)


Call:
lm(formula = Sales ~ ., data = CigaretteConsumption[-c(9, 29,
    30), -1])

Residuals:
    Min      1Q  Median      3Q     Max
-40.155  -8.663  -2.194   6.301  36.043

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.68317  136.24526   0.739 0.464126
Age           1.84871    1.79266   1.031 0.308462
HS           -1.17246    0.52712  -2.224 0.031696 *
Income        0.02084    0.00546   3.817 0.000448 ***
Black        -0.30346    0.40567  -0.748 0.458702
Female        1.12460    3.07908   0.365 0.716810
Price        -2.78195    0.57818  -4.812 2.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.02 on 41 degrees of freedom
Multiple R-squared:  0.4871,   Adjusted R-squared:  0.412
F-statistic: 6.489 on 6 and 41 DF,  p-value: 7.195e-05
```

models. There is also significant reduction in the residual standard error (from 28.17 to 15.02). Furthermore, R^2 has improved from 0.32 to 0.49.

This is not to say that we are done with the analysis. There are more checks, such as checks for normality, heteroscedasticity, etc., that are pending.

3.13 Collinearity

Let us revisit Example 2.3 (cars) which we analyzed in Sects. 3.7 and 3.10. We incorporate data on an additional variable, namely, the weights (WT) of these cars. A linear regression of MPG on HP, HP_SQ, VOL, and WT is performed. The output is given below.

Compare the output in Tables 7.5 and 7.9. The following points emerge:

(a) WT is insignificant, as noticed in Table 7.9.
(b) VOL which is highly significant in Table 7.5 turns out to be highly insignificant in Table 7.9. Thus, once we introduce WT, both VOL and WT become insignificant which looks very surprising.
(c) The coefficient estimates of VOL in Tables 7.5 and 7.9 (corresponding to the models without and with WT, respectively) are -0.1018 and -0.0049 which are

Table 7.9 Output for the regression of MPG on HP, HP_SQ, VOL, and WT

```
> model5<-lm(MPG ~ HP + HP_sq + VOL + WT, data = Cars)
> summary(model5)

Call:
lm(formula = MPG ~ HP + HP_sq + VOL + WT, data = cars)

Residuals:
    Min      1Q    Median      3Q       Max
-8.3218  -2.0723  0.5592     1.7386   10.9699

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.734e+01  2.126e+00  36.374  < 2e-16 ***
HP          -4.121e-01  3.099e-02 -13.299  < 2e-16 ***
HP_sq        9.053e-04  9.105e-05   9.943 2.13e-15 ***
VOL         -4.881e-02  3.896e-01  -0.125    0.901
WT          -1.573e-01  1.156e+00  -0.136    0.892
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.078 on 76 degrees of freedom
Multiple R-squared:  0.8921,   Adjusted R-squared:  0.8864
F-statistic:   157 on 4 and 76 DF, p-value: < 2.2e-16
```

quite far apart. Also, the corresponding standard errors are 0.01795 and 0.3896. Once we introduce WT, the standard error of the coefficient for VOL changes by more than 20 times. Furthermore, the value of the coefficient is halved. Thus, we notice that the standard error has gone up by as high as 20 times and the magnitude of the coefficient is halved once we introduce the variable WT.

(d) There is virtually no change in R square.

Since there is virtually no change in R square, it is understandable why WT is insignificant. But why did VOL, which was highly significant before WT was introduced, became highly insignificant once WT is introduced? Let us explore. Let us look at the scatterplot matrix (Fig. 7.11).

One thing that is striking is that VOL and WT are almost perfectly linearly related. So in the presence of WT, VOL has virtually no additional explaining capacity for the variation in the residual part of MPG not already explained by WT. The same is the situation with WT that it has no additional explaining capacity in the presence of VOL. If both of them are in the list of regressors, both of them become insignificant for this reason. Let us look at the added variable plots for VOL and WT in the model corresponding to Table 7.9 which confirm the same thing (Fig. 7.12).

It is said that there is a collinear relationship among some of the regressors if one of them has an almost perfect linear relationship with others. If there are collinear relationships among regressors, then we say that there is the problem of *Collinearity*.

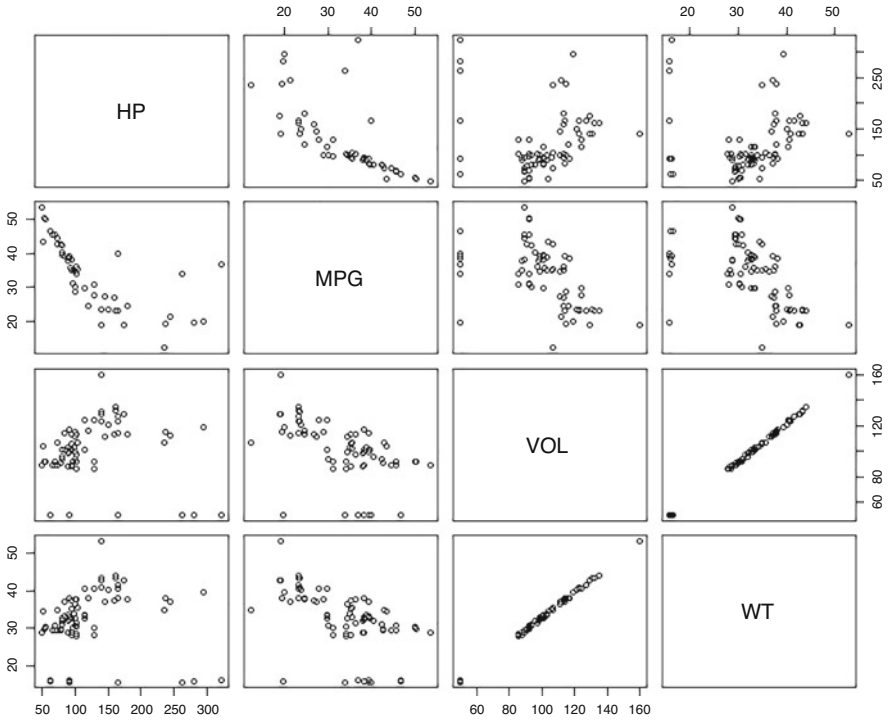


Fig. 7.11 Scatterplot matrix with variables MPG, HP, VOL, and WT

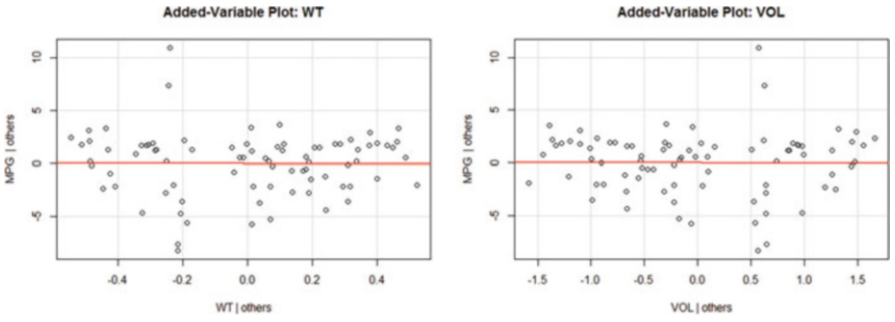


Fig. 7.12 Added variable plots for WT and VOL

Why should we care if there is collinearity? What are the symptoms? How do we detect collinearity and if detected, what remedial measures can be taken?

Some of the symptoms of collinearity are as follows:

(a) R square is high, but almost all the regressors are insignificant.
(b) Standard errors of important regressors are large, and important regressors become insignificant.

(c) Very small changes in the data produce large changes in the regression coefficient estimates.

We noticed (b) above in our MPG example.

How does one detect collinearity? If the collinear relation is between a pair of regressors, it can be detected using the scatterplot matrix and the correlation matrix. In the MPG example, we detected collinearity between VOL and WT this way. Suppose that more than two variables are involved in a collinear relationship. In order to check whether a regressor is involved in a collinear relationship, one can use *variance inflation factors (VIF)*.

The variance inflation factor for the i^{th} regressor, denoted by VIF_i is defined as the factor by which the variance of a single observation, σ^2, is multiplied to get the variance of the regression coefficient estimate of the i^{th} regressor. It can be shown that VIF_i is the reciprocal of $1 - R_i^2$ where R_i^2 is the coefficient of multiple determination of the i^{th} regressor with the other regressors. The i^{th} regressor is said to be involved in a collinear relationship if R_i^2 is large. There is no unanimity on how large is considered to be large, but 95% is an accepted norm. If R_i^2 is 95% or larger, then VIF_i is at least 20.

Variance decomposition proportions (VP): Table 7.10 (see Belsley et al. 2005) is sometimes used to identify the regressors involved in collinear relationships. We shall explain below how to use the VP table operationally for the case where there are four regressors and an intercept. The general case will follow similarly. For more details and the theoretical basis, one can refer to BKW.

By construction, the sum of all the elements in each column starting from column 3 (columns corresponding to the intercept and the regressors), namely, $\sum_{j=0}^{4} \pi_{ji}$, is 1 for $i = 0,1,\ldots,4$.

Algorithm

Step 0: Set $i = 0$.

Step 1: Check whether the condition index c_{5-i} is not more than 30. If yes, declare that there is no serious issue of collinearity to be dealt with and stop. If no, go to step 2.

Step 2: Note down the variables for which the π values in the $(5 - i)^{th}$ row are at least 0.5. Declare these variables as variables involved in a collinear relationship. Go to Step 3.

Table 7.10 Variance decomposition proportions table

S. No.	Condition index	Intercept	X_1	X_2	X_3	X_4
1	c_1	π_{00}	π_{01}	π_{02}	π_{03}	π_{04}
2	c_2	π_{10}	π_{11}	π_{12}	π_{13}	π_{14}
3	c_3	π_{20}	π_{21}	π_{22}	π_{23}	π_{24}
4	c_4	π_{30}	π_{31}	π_{32}	π_{33}	π_{34}
5	c_5	π_{40}	π_{41}	π_{42}	π_{43}	π_{44}

Table 7.11 Variance inflation factors

```
> vif(model5)
      HP      HP_sq        VOL         WT
26.45225   26.32768  637.51477  634.06751
```

Table 7.12 Variance decomposition proportions

```
> colldiag(model5)
Condition
Index            Variance Decomposition Proportions
             intercept       HP          HP_sq   VOL     WT
1    1.000       0.001    0.000          0.001   0.000   0.000
2    2.819       0.004    0.001          0.024   0.000   0.000
3   11.652       0.511    0.002          0.008   0.000   0.000
4   29.627       0.358    0.983          0.956   0.000   0.000
5  338.128       0.127    0.013          0.012   1.000   0.999
```

Step 3: Delete the row $5 - i$ from the table and calibrate the π values in each column corresponding to the intercept and the regressors so that the corresponding columns is 1.

Step 4: Replace i by $i+1$ and go to step 1,

When the algorithm comes to a stop, say, at $i = 3$, you have 2 (i.e., $(i - 1)$) collinear relationships with you.

Let us return to the MPG example and the model that led us to the output in Table 13.1. The VIFs and the variance decomposition proportions table are given in Tables 7.11 and 7.12:

The VIFs of VOL and WT are very high (our cutoff value is about 20), and thereby imply that each of VOL and WT is involved in collinear relationships. This also explains what we already observed, namely, VOL and WT became insignificant (due to large standard errors). The VIFs of HP and HP_SQ are marginally higher than the cutoff.

From the variance decompositions proportions table, we see that there is a collinear relationship between VOL and WT (condition index is 338.128 and the relevant π values corresponding to VOL and WT are 1 and 0.999, respectively). The next largest condition index is 29.667 which is just about 30. Hence, we can conclude that we have only one mode of collinearity.

What remedial measures can be taken once the collinear relationships are discovered?

Let us start with the easiest. If the intercept is involved in a collinear relationship, subtract an easily interpretable value close to the mean of each of the other regressors involved in that relationship and run the regression again. You will notice that the regression coefficient estimates and their standard errors remain the same as in the earlier regression. Only the intercept coefficient and its standard error will change. The intercept will no longer be involved in the collinear relationship.

Consider one collinear relationship involving some regressors. One can delete the regressor that has the smallest partial correlation with the response variable given

the other regressors. This takes care of this collinear relationship. One can repeat this procedure with the other collinear relationships.

We describe below a few other procedures, stepwise regression, best subset regression, ridge regression, and lasso regression, which are commonly employed to combat collinearity in a blanket manner.

It may be noted that subset selection in the pursuit of an appropriate model is of independent interest for various reasons, some of which we mention below.

(a) The "kitchen sink" approach of keeping many regressors may lead to collinear relationships.
(b) Cost can be a consideration, and each regressor may add to the cost. Some balancing may be needed.
(c) Ideally there should be at least ten observations per estimated parameter. Otherwise one may find significances by chance. When the number of observations is not large, one has to restrict the number of regressors also.

The criteria that we describe below for selecting a good subset are based on the residual sum of squares. Let us assume that we have one response variable and k regressors in our regression problem. Suppose we have already included r regressors ($r < k$) into the model. We now want to introduce one more regressor from among the remaining $k - r$ regressors into the model. The following criteria place a penalty for bringing in a new regressor. The coefficient of multiple determination, R^2 unfortunately never decreases when a new regressor is introduced. However, adjusted R^2 can decrease when a new regressor is introduced if it is not sufficiently valuable. We introduce a few other criteria here for which the value of the criterion increases unless the residual sum of squares decreases sufficiently by introducing the new regressor, indicating that it is not worth introducing the regressor under consideration. The current trend in stepwise regression is to start with the model in which all the k regressors are introduced into the model and drop the regressors as long as the criterion value decreases and stop at a stage where dropping a regressor increases the criterion value. The object is to get to a subset of regressors for which the criterion has the least value. We use the following notation:

$N =$ The number of observations
$k =$ The total number of regressors
$r =$ The number of regressors used in the current model
$\sum_{i=1}^{N}(y_i - \overline{y})^2 =$ The sum of squared deviations of the response variable from its mean
$\left(R_0^2\right)_r =$ The sum of squared residuals when the specific subset of r regressors is used in the model
$\left(R_0^2\right)_k =$ The sum of squared residuals when all the k regressors are used in the model

The major criteria used in this connection are given below:

(a) Adjusted R^2: $1 - \frac{N-1}{N-r}\frac{\left(R_0^2\right)_r}{\sum_{i=1}^{N}(y_i-\overline{y})^2}$ (see also Eq. 7.11)

(b) AIC : $\log\left(\frac{(R_0^2)_r}{N}\right) + \frac{2r}{N}$

(c) BIC : $\log\left(\frac{(R_0^2)_r}{N}\right) + \frac{r \log N}{N}$

(d) Mallow's C_p: $\frac{(R_0^2)_r}{(R_0^2)_k/(Nk-1)} - n + 2r$

Note: The AIC, or Akaike Information Criterion, equals twice the negative of the log-likelihood penalized by twice the number of regressors. This criterion has general applicability in model selection. The BIC or Bayes Information Criterion is similar but has a larger penalty than AIC and like AIC has wider application than regression.

Among (a), (b), and (c) above, the penalty for introducing a new regressor is in the ascending order. The criterion (d) compares the residual sum of squares of the reduced model with that of the full model. One considers the subset models for which C_p is close to r and chooses the model with the least number of regressors from among these models.

We illustrate the stepwise procedure with the cigarette consumption example (Example 7.1) using the criterion AIC. We give the R output as in Table 7.13.

The first model includes all the six regressors, and the corresponding AIC is 266.56. In the next stage one of the six regressors is dropped at a time, keeping all other regressors, and the AIC value is noted. When Female is dropped, keeping all other regressors intact, the AIC is 264.71. Likewise, when age is dropped, the AIC is 265.79, and so on. We notice the least AIC corresponds to the model dropping Female. In the next stage, the model with the remaining five regressors is considered. Again, the procedure of dropping one regressor from this model is considered and the corresponding AIC is noted. (The dropped regressor is brought back and its AIC is also noted. In the case of AIC this is not necessary because this AIC value is already available in a previous model. However, for the case of Adj. R^2 this need not necessarily be the case.) We find that the least AIC equaling 263.23 now corresponds to the model which drops Black from the current model with the five regressors. The procedure is repeated with the model with the four regressors. In this case dropping any variable from this model yields an increased AIC. The stepwise procedure stops here. Thus, the stepwise method yields the model with the four regressors, age, HS, income, and price. The corresponding estimated model is given below.

Compare Tables 7.14 and 7.8. We notice that the significance levels of HS, Income, and Price have remained the same (in fact, the p-values are slightly smaller in the subset model). Age, which was insignificant in the full model (Table 7.8), is now (Table 7.14) significant at 5% level (p-value is 0.039). So when some undue regressors are dropped, some of the insignificant regressors may become significant. It may also be noted that while R^2 dropped marginally from 0.4871 to 0. 4799 corresponding to full model and the subset model, respectively, there is a substantial increase in adjusted R^2 from 0. 412 in the full model to 0.4315 in the subset model.

Table 7.13 Stepwise regression

```
> step <-stepAIC (model1, direction="both")

Start:  AIC=266.56
Sales ~ Age + HS + Income + Black + Female + Price

          Df Sum of Sq     RSS    AIC
- Female  1      30.1  9284.4 264.71
- Black   1     126.3  9380.6 265.21
- Age     1     240.0  9494.3 265.79
<none>                 9254.3 266.56
- HS      1    1116.7 10370.9 270.03
- Income  1    3288.6 12542.8 279.15
- Price   1    5225.4 14479.7 286.05

Step:  AIC=264.71
Sales ~ Age + HS + Income + Black + Price

          Df Sum of Sq     RSS    AIC
- Black   1      99.4  9383.8 263.23
<none>                 9284.4 264.71
- Age     1     629.1  9913.4 265.86
+ Female  1      30.1  9254.3 266.56
- HS      1    1099.8 10384.1 268.09
- Income  1    3366.8 12651.2 277.57
- Price   1    5198.8 14483.2 284.06

Step:  AIC=263.23
Sales ~ Age + HS + Income + Price

          Df Sum of Sq     RSS    AIC
<none>                 9383.8 263.23
+ Black   1      99.4  9284.4 264.71
+ Female  1       3.2  9380.6 265.21
- Age     1     991.8 10375.6 266.05
- HS      1    1573.0 10956.8 268.67
- Income  1    3337.8 12721.6 275.83
- Price   1    5216.3 14600.1 282.44
```

Best Subset Regression

In stepwise regression, we considered 15 subsets, as can be seen from Table 7.14. But there are $2^6 - 1 = 63$ subsets with at least one regressor. The Best Subset regression (using AIC or BIC) considers all these subsets in a systematic manner and delivers that subset for which the AIC (or BIC) is the least. In case of Adjusted R^2, it delivers that subset for which the adjusted R^2 is the largest. However, if the number of regressors is large, it generally gets unwieldy to search for the best subset.

Table 7.14 Best Stepwise model from stepAIC

```
> best_model_step <-lm(Sales ~ Age + HS + Income + Price, data=Cig_data)
> summary(best_model_step)

Call:
lm(formula = Sales ~ Age + HS + Income + Price, data = Cig_data)

Residuals:
    Min      1Q  Median      3Q     Max
-40.196  -8.968  -1.563   8.525  36.117

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.966923  37.849699   3.302 0.001940 **
Age           2.628064   1.232730   2.132 0.038768 *
HS           -0.894433   0.333147  -2.685 0.010267 *
Income        0.019223   0.004915   3.911 0.000322 ***
Price        -2.775861   0.567766  -4.889 1.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.77 on 43 degrees of freedom
Multiple R-squared:  0.4799,   Adjusted R-squared:  0.4315
F-statistic:  9.92 on 4 and 43 DF,  p-value: 8.901e-06
```

In such a case, stepwise regression can be employed. For the cigarette consumption data, the best subset regression using AIC leads to the same regressors as in the stepwise regression using AIC.

Ridge Regression and Lasso Regression

In the least squares method (see estimation of parameters in Sects. 3.6 and 3.7), we estimate the parameters by minimizing the error sum of squares. Ridge regression and lasso regression minimize the error sum of squares subject to constraints that place an upper bound on the magnitude of the regression coefficients. Ridge regression minimizes the error sum of squares subject to $\sum \beta_i^2 \leq c$, where c is a constant. Lasso regression (Tibshirani 1996) minimizes the error sum of squares subject to $\sum \mid \beta_i \mid \leq d$, where d is a constant. Both these methods are based on the idea that the regression coefficients are bounded in practice. In ridge regression all the regressors are included in the regression and the coefficient estimates are nonzero for all the regressors. However, in lasso regression it is possible that some regressors may get omitted. It may be noted that both these methods yield biased estimators which have some interesting optimal properties under certain conditions.

The estimates for the regression coefficients for the Cigarette consumption data are given in Table 7.15.

In ridge regression, all the regressors have nonzero coefficient estimates which are quite different from those obtained in Table 7.8 or Table 7.14. The signs match, however. Lasso regression drops the same variables as in stepwise regression and best subset regression. The coefficient estimates are also not too far off.

In practice, it is better to consider stepwise/best subset regression and lasso regression and compare the results before a final subset is selected.

Table 7.15 Coefficients using ridge and lasso regressions

Ridge regression

```
> Cig_data<-CigaretteConsumption[-c(9,29,30),-1]
> lambda <- 10^seq(10, -2, length = 100)
> ridge.mod <- glmnet(as.matrix(Cig_data[,-7]),as.matrix(Cig_data[,7]), alpha
 = 0, lambda = lambda)
> predict(ridge.mod, s = 0, Cig_data[,-7], type = 'coefficients')[1:6,]

 (Intercept)          Age           HS       Income        Black       Female

100.59166959   1.85775082  -1.16705566   0.02079292  -0.30019556   1.11722448


> cv.out <- cv.glmnet(as.matrix(Cig_data[,-7]),as.matrix(Cig_data[,7]), alpha
 = 0)
> bestlam <- 13.1
> coef(cv.out, s=bestlam)

7 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 62.407535361
Age          1.602311458
HS          -0.258625440
Income       0.007815691
Black        0.038966928
Female       0.892841343
Price       -1.366561732
```

Lasso regression

```
> lasso.mod <- glmnet(as.matrix(Cig_data[,-7]),as.matrix(Cig_data[,7]), alpha
 = 1, lambda = lambda)
> predict(lasso.mod, s = 0, Cig_data[,-7], type = 'coefficients')[1:6,]

 (Intercept)          Age           HS       Income        Black       Female

101.17866407   1.86717911  -1.16229575   0.02075092  -0.29588646   1.09634187


> cv.out1 <- cv.glmnet(as.matrix(Cig_data[,-7]),as.matrix(Cig_data[,7]), alph
a = 1)
> bestlam1 <- 0.7
> coef(cv.out1, s=bestlam1)

7 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 120.81017762
Age           2.40000527
HS           -0.70182973
Income        0.01614734
Black         .
Female        .
Price        -2.46939620
```

3.14 Dummy Variables

Gender discrimination in wages is a highly debated topic. Does a man having the same educational qualification as a woman earn a higher wage on average? In order to study the effect of gender on age controlling for the educational level, we use data

Table 7.16 Wage equation for males and females

```
> dummy_reg <-lm(Wage ~ Education + Female, data=female)
> summary(dummy_reg)

Call:
lm(formula = Wage ~ Education + Female, data = female)

Residuals:
    Min      1Q  Median      3Q     Max
-12.440  -3.603  -1.353   1.897  91.603

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.85760    0.51279  -1.672   0.0945 .
Education    0.95613    0.04045  23.640   <2e-16 ***
Female      -2.26291    0.15198 -14.889   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.006 on 8543 degrees of freedom
Multiple R-squared:  0.07949,  Adjusted R-squared:  0.07928
F-statistic: 368.9 on 2 and 8543 DF,  p-value: < 2.2e-16
```

on hourly average wage, years of education, and gender on 8546 adults collected by Current Population Survey (1994), USA. Gender is a qualitative attribute. How does one estimate the effect of gender on hourly wages? An individual can either be male or female with no quantitative dimension. We need a quantifiable variable that can be incorporated in the multiple regression framework, indicating gender. One way to "quantify" such attributes is by constructing an artificial variable that takes on values 1 or 0, indicating the presence or absence of the attribute. We can use 1 to denote that the person is a female and 0 to represent a male. Such a variable is called a Dummy Variable. A *Dummy Variable* is an indicator variable that reveals (indicates) whether an observation possesses a certain characteristic or not. In other words, it is a *device to classify data into mutually exclusive categories such as male and female.*

We create the dummy variable called "female," where female = 1, if gender is female and female = 0, if gender is male. Let us write our regression equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 \text{female} + \varepsilon,$$

where Y = Wage and X = Years of education.

The regression output is shown in Table 7.16.

Education is significant. One year of additional schooling will lead to an increase in wage by $0.956. How do we interpret the intercept and the coefficient of Female? The estimated regression equation is

$$\widehat{Y} = -0.8576 + 0.9561 \text{ Education} - 2.2629 \text{ Female}.$$

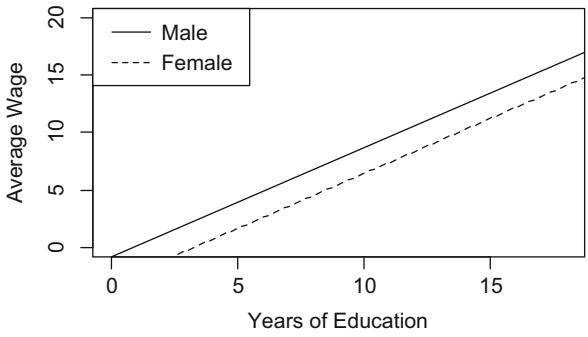


Fig. 7.13 Parallel regression lines for the wages of males and females

When Female = 1 (that is for females), intercept = $\beta_0 + \beta_2$. When Female = 0 (that is for males), intercept = β_0.

Thus, for the same level of education, the difference in the average wage between a female and a male is the difference in intercept, β_2 (−2.2629). In other words, a female receives $2.2629 less average hourly wage as compared to a male. A male with 10 years of education earns $(−0.8576 + 0.9561 * 10) = $8.7034 of hourly wage. A female with 10 years of education earns $(−0.8576 + 0.9561 * 10-2.2629) = $6.4405 of hourly wage. Thus $\beta_2 = -2.2629$ is the additional effect of female on wages. Here, male is called the *base category* because the effect of the gender female on wages is measured over and above that of being male.

Here, the dummy variable acts as an Intercept Shifter. Notice that the regression line for the male is −0.8576 + 0.9561 * Education and that for female is −3.1205 + 0.9561 * Education. Thus, the two regression lines differing only by intercept are parallel shifts of each other. Such a dummy is sometimes called intercept dummy (Fig. 7.13).

At this point one may wonder: what if I take the dummy Male = 1 for male and 0 for female? Will the regression results for the same data set change? No, they will remain the same.

Intercept Dummy fits a different intercept for each qualitative characteristic. What if the relationship is different—the effect of female is not just a uniform negative number added to the base wage but instead depends on the level of education? In fact, discrimination may work in many ways: a uniform lower wage for women across all education levels (modeled by intercept dummy), or lower wages for women versus men as education increases (incremental return from education is more for men than for women) or the gap in the wages of men and women may reduce with more education. Clearly, the first model specification is inadequate if the second or the third case occurs! This leads us to a new specification.

Relax the assumption of parallel slopes or equal returns to education for men and women by introduction of new a variable "interaction term" defined as:

$$\text{Education}^{*}\text{Female} = \begin{cases} Education & Female = 1 \\ 0 & Female = 0 \end{cases}.$$

Table 7.17 Wage equation with interaction term between female and education

```
> dummy_reg2 <-lm(Wage ~ Education + Female + Female.Education, data=female)
> summary(dummy_reg2)

Call:
lm(formula = Wage ~ Education + Female + Female.Education, data = female)

Residuals:
    Min      1Q  Median      3Q     Max
-12.168  -3.630  -1.340   1.904  91.743

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.07569    0.69543   0.109    0.913
Education         0.88080    0.05544  15.887  < 2e-16 ***
Female           -4.27827    1.02592  -4.170 3.07e-05 ***
Female.Education  0.16098    0.08104   1.986    0.047 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.005 on 8542 degrees of freedom
Multiple R-squared:  0.07992,  Adjusted R-squared:  0.07959
F-statistic: 247.3 on 3 and 8542 DF,  p-value: < 2.2e-16
```

Thus, the interaction term is the product of the dummy variable and education. Female and Education interact to produce a new variable Female * Education. For the returns to education data, here is the regression output when we include the interaction term (Table 7.17).

We notice that the interaction effect is positive and significant at the 5% level. What does this mean?

The predicted wage for Male is given by

$$\begin{aligned} &0.07569 + 0.88080^{*}\text{Education} - 4.27827^{*}\ (\text{Female}{=}0) + 0.16098^{*} \\ &\{(\text{Female}{=}0) * \text{Education}\} \\ &\quad = 0.07569 + 0.88080^{*}\text{Education} \end{aligned}$$

The predicted Wage for Female is given by

$$\begin{aligned} &0.07569 + 0.88080^{*}\text{Education} - 4.27827^{*}\ (\text{Female} = 1) + 0.16098^{*} \\ &\{(\text{Female} = 1) * \text{Education}\} \\ &\quad = 0.07569 + 0.88080^{*}\text{Education} - 4.27827 + 0.16098^{*}\text{Education} \\ &\quad = -4.20258 + 1.04178^{*}\text{Education} \end{aligned}$$

Notice that for the two regression equations, both the slope and the intercepts are different!

Good news for the feminist school! An additional year of education is worth more to females because $\beta_3 = 0.16098 > 0$. An additional year of education is worth about \$ 0.88 extra hourly wage for men. An additional year of education is worth about \$1.04178 extra hourly wage for women. A man with 10 years of education earns: \$ (0.07569 + 0.88080 * 10) = \$ 8.88369 average hourly wage. A woman with 10 years of education earns: \$ (−4.20258 + 1.04178 * 10) = \$ 6.21522 average hourly wage.

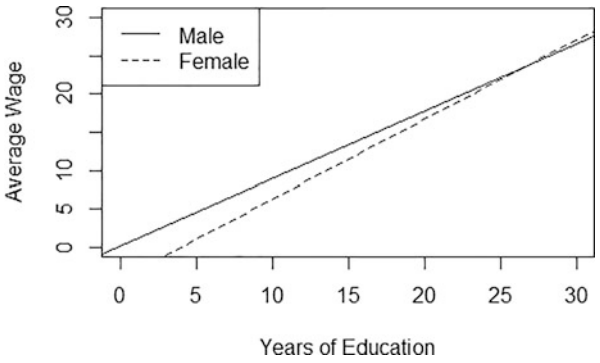


Fig. 7.14 Regression equations for males and females

Thus, we see that there are two effects in work: (a) The female wage-dampening effect (through lower intercept for women) across education and (b) narrowing of gap in wage with years of education. This is depicted visually in Fig. 7.14.

It appears from the above figure that women start earning more than men starting from 27 years of education. Unfortunately, this conclusion cannot be drawn from the data on hand as the maximum level of education in our data is 18 years.

So far we have considered the case of two categories. In the returns to education data set considered in this section, the dataset refer to the variable "PERACE." An individual can come from five different races—WHITE, BLACK, AMERICAN INDIAN, ASIAN, OR PACIFIC ISLANDER, OTHER. The question under consideration is: Is there also racial discrimination in wages? How to model race as a regressor in the wage determination equation? Clearly, one dummy variable taking values 0, 1 will not work! One possibility is that we assign five different dummy variables for the five races. $D_1 = 1$, if white and $= 0$, otherwise; $D_2 = 1$, if Black and $= 0$, otherwise; $D_3 = 1$, if American Indian and $= 0$, otherwise; $D_4 = 1$, if Asian or Pacific Islander and $= 0$, otherwise; and $D_5 = 1$, if other and $= 0$, otherwise.

The regression output after introducing these dummies into the model is as given in Table 7.18.

What are the NAs corresponding to the others? Clearly, there is some problem with our method. R did not compute estimates for the "Other" dummy citing the reason "1 not defined because of singularities." The issue actually is the following: for every individual one and only one D1–D5 is 1 and the rest are all 0. Hence the sum of D1–D5 is 1 for each individual. Hence there is perfect collinearity between the intercept and the dummies D1–D5.

What is the solution? When you have n categories, assign either (a) n Dummies and no intercept OR b) $(n - 1)$ Dummies and an intercept. In our example, for the five races, either assign four dummies and an intercept or just assign five dummies but no intercept.

R automatically inputted four dummies to denote the five race categories and one dummy to denote gender. If female $= 0$, and all four race dummies (D1, D2, D3, D4) $= 0$, then estimated regression equation is

Table 7.18 Wage equation for different ethnic groups

```
> dummy_reg3 <-lm(Wage ~ Education + Female + D1+D2+D3+D4+D5, data=female)
> summary(dummy_reg3)

Call:
lm(formula = Wage ~ Education + Female + D1 + D2 + D3 + D4 + D5, data = femal
e)

Residuals:
    Min      1Q  Median      3Q     Max
-12.527  -3.615  -1.415   1.963  92.124

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.04604    0.87495  -2.338   0.0194 *
Education    0.95173    0.04055  23.470   <2e-16 ***
Female      -2.25709    0.15205 -14.845   <2e-16 ***
D1           1.34520    0.74842   1.797   0.0723 .
D2           0.71014    0.77995   0.911   0.3626
D3           1.01328    0.98776   1.026   0.3050
D4           0.72720    0.85495   0.851   0.3950
D5                NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.003 on 8539 degrees of freedom
Multiple R-squared:  0.08069,  Adjusted R-squared:  0.08004
F-statistic: 124.9 on 6 and 8539 DF,  p-value: < 2.2e-16
```

Table 7.19 Transforming education into categories

Education years	Category
≤12	School
13–16	College
≥17	Higher education

$$\widehat{\mathbf{y}} = -2.04604 + 0.95173\,{}^{*}\text{Education}$$

Thus, the intercept here denotes the effect of all the excluded categories—that is, the effect of the base category "Other male." All the dummies measure the effect over and above the base category "Other male." Looking at the R-output in Table 7.18, we infer that none of the race dummies is significant. (White is just about significant at the 10% level.) Whether you are white or black or any other race does not affect your wages. No racial discrimination in wages! But since the coefficient female is negative in estimate and highly significant, there is gender discrimination in wages!

Consider the "Education" variable. Till now, we were estimating the incremental effect of an additional year of education on wages. Moving education level from class 5 to class 6 is not so much likely to make a difference to wages. Rather, going from school to college or college to higher education may make a difference. Perhaps a more sensible way to model education is to group it into categories as show in Table 7.19.

The categories defined here (school, college, and higher education) are also qualitative but they involve an ordering—a college graduate is higher up the

Table 7.20 Transforming an ordinal variable into dummy variables

Observation	Category	College (D1)	Higher_Ed (D2)
1	School	0	0
2	College	1	0
3	Higher education	1	1

Table 7.21 Wage equation with education dummies

```
> dummy_reg4 <-lm(Wage ~ Female + College + Higher_Ed + Female.College + Fema
le.Higher_Ed, data=female)
> summary(dummy_reg4)

Call:
lm(formula = Wage ~ Female + College + Higher_Ed + Female.College +
    Female.Higher_Ed, data = female)

Residuals:
    Min      1Q  Median      3Q     Max
-13.694  -3.688  -1.539   2.068  91.112

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      10.1322     0.1440  70.356  < 2e-16 ***
Female           -2.2438     0.2049 -10.950  < 2e-16 ***
College           1.7267     0.2281   7.570 4.12e-14 ***
Higher_Ed         6.0848     0.7416   8.205 2.65e-16 ***
Female.College    0.3164     0.3144   1.006    0.314
Female.Higher_Ed  0.6511     1.0940   0.595    0.552
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.093 on 8540 degrees of freedom
Multiple R-squared:  0.05676,  Adjusted R-squared:  0.05621
F-statistic: 102.8 on 5 and 8540 DF,  p-value: < 2.2e-16
```

education ladder than a school pass out and one with a degree higher than a college degree is still higher up. To incorporate the effect of education categories, ordinary dummy variable is not enough. The effect of College on wages is over and above the effect of schooling on wages. The effect of "Higher Education" on wages will be some notches higher than that of college education on wages. Assign dummy variables as shown in Table 7.20.

The output after incorporating these dummies usually called *ordinal dummies* is shown in Table 7.21.

How do we interpret this output? The increment in the hourly wage for completing college education over high school is $1.7267 and that for completing higher education over college degree is $6.0848. Both these are highly significant (based on the p-values).

For identifying that there is a dummy in play using residual plots see Exercise 7.4.

For an interesting application of dummy variables and interactions among them, see Exercise 7.5.

For the use of interaction between two continuous variables in linear regression, see the chapter on marketing analytics (Chap. 19).

3.15 Normality and Transformation on the Response Variable

One of the assumptions we made in the linear regression model (Sect. 3.5) is that the errors are normally distributed. Do the data on hand and the model proposed support this assumption? How does one check this? As we mentioned several times in Sect. 3.7 and later, the residuals of different types are the representatives of the errors. We describe below one visual way of examining the normality of errors, known as Q-Q plot of the studentized residuals (see 7.43) for the definition of a studentized residual). In Q-Q plot, Q stands for quantile. First, order the observations of a variable of interest in the ascending order. We recall that the first quartile is that value of the variable below which 25% of the ordered observations lie and above which 75% of the ordered observations lie. Let p be a fraction such that $0 < p < 1$. Then the p^{th} quantile is defined as that value of the variable below which a proportion p of the ordered observations lie and above which a proportion $1 - \text{p}$ of the ordered observations lie. In the normal Q-Q plot of the studentized residuals, the quantiles of the studentized residuals are plotted against the corresponding quantiles of the standard normal distribution. This plot is called the normal Q-Q plot of the studentized residuals. Let the i^{th} quantile of studentized residuals (often referred to as sample quantile) be denoted by q_i and the corresponding quantile of the standard normal distribution (often referred to as theoretical quantile) be denoted by t_i. If the normality assumption holds, in the ideal case, $(t_i, q_i), i = 1, \ldots, N$ fall on a straight line. Since we are dealing with a sample, it is not feasible that all the points fall on a straight line. So a confidence band around the ideal straight line is also plotted. If the points go out of the band, then there is a concern regarding normality. For more details regarding the Q-Q plots, you can read stackexchange.[2]

We give below the Q-Q plots of the studentized residuals for different models related to Example 2.1. For Example 2.1, we consider the models AT vs WC, $\sqrt{AT}$ vs WC, and log AT vs WC and WC^2. (We shall explain in the case study later why the latter two models are of importance.) The Q_Q plots are given in Fig. 7.15.

Compare the three plots in Fig. 7.15. We find that AT vs WC is not satisfactory as several points are outside the band. The plot of $\sqrt{AT} vs WC$ is somewhat better and that of $\log AT vs WC, WC^2$ is highly satisfactory.

There are also formal tests for testing for normality. We shall not describe them here. The interested reader may consult Thode (2002). A test tells us whether the null hypothesis of normality of errors is rejected. However, the plots are helpful in taking the remedial measures, if needed.

If the residuals support normality, then do not make any transformation as any nonlinear transformation of a normal distribution never yields a normal distribution. On the other hand, if Q-Q plot shows significant deviation from normality, it may be due to one or more of several factors, some of which are given below:

[2] https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot (Accessed on Feb 5, 2018).

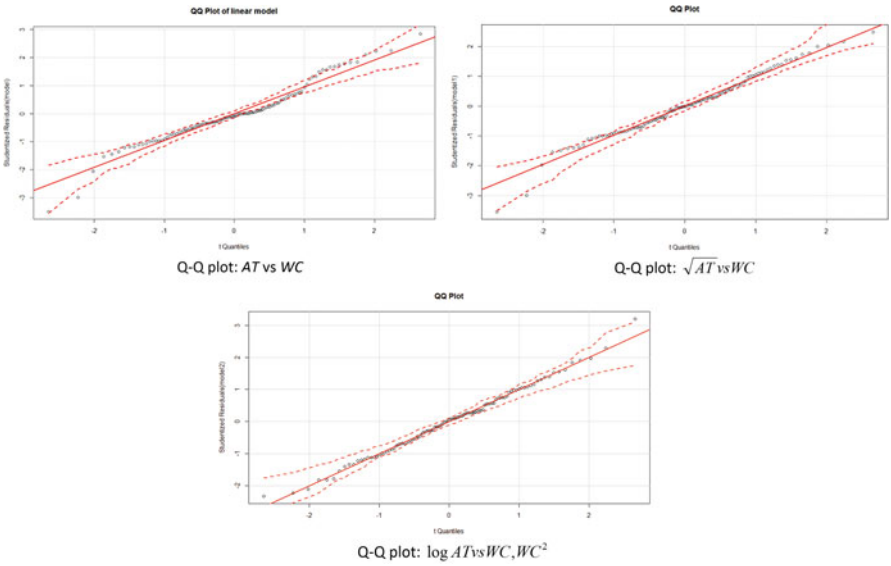


Fig. 7.15 Q-Q plots for the adipose tissue example

(a) Presence of influential observations.
(b) Omission of some important regressors which are correlated with the regressors included in the model.
(c) The response variable requires a transformation.

When you find nonnormality of studentized residuals, first check (a) and (b) above and if one of them is the case, then take care of them by the techniques we already developed. If the nonlinearity still persists, then contemplate a transformation on the response variable. Power transformations are what are commonly used. There is an oriented way of determining the power popularly known as the Box–Cox transformation. We describe the same hereunder.

Box–Cox Transformation

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda \tilde{Y}^{\lambda-1}} & \lambda \neq 0 \\ \tilde{Y} \log Y & \lambda = 0 \end{cases} \tag{7.45}$$

where $\tilde{Y} = (y_1 \times y_2 \times \ldots y_N)^{\frac{1}{N}}$ is the geometric mean of the data on the response variable.

With this transformation, the model will now be $Y^{(\lambda)} = Z\beta + \varepsilon$. This model has an additional parameter λ over and above the parametric vector β. Here the errors are minimized over β and λ by the method of least squares. In practice, for various values of λ in $(-2, 2)$, the parametric vector β is estimated and the corresponding estimated log-likelihood is computed. The values of the estimated log-likelihood (y-

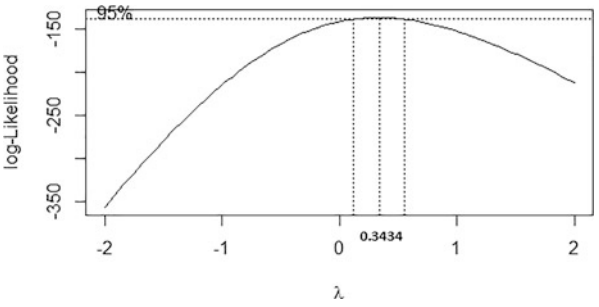


Fig. 7.16 Box–Cox plot

axis) are plotted against the corresponding value of λ. The value of λ at which the estimated log-likelihood is maximum is used in (7.45) to compute the transformed response variable. Since the value of λ is estimated from a sample, a confidence interval for λ is also obtained. In practice, that value of λ in the confidence interval is selected which is easily interpretable. We shall illustrate this with Example 2.1 where the response variable is chosen as *AT* and the regressor as *WC*. The plot of log-likelihood against λ is given below (Fig. 7.16).

Notice that the value of λ at which the log-likelihood is the maximum is 0.3434. This is close to $1/3$. Still it is not easily interpretable. But 0.5 is also in the confidence interval and this corresponds to the square-root which is more easily interpretable. One can use this and make the square-root transformation on the *AT* variable.

3.16 Heteroscedasticity

One of the assumptions in the least squares estimation and testing was that of equal variance of the errors (E in LINE). If this assumption is violated, then the errors do not have equal status and the standard least squares method is not quite appropriate. The unequal variance situation is called heteroscedasticity. We shall talk about the sources for heteroscedasticity, the methods for detecting the same, and finally the remedial measures that are available.

Consider AT–Waist problem (Example 2.1). Look at Fig. 7.2 and observation (ii) following the figure. We noticed that the variation in adipose tissue area increases with increasing waist circumference. This is a typical case of heteroscedasticity.

We shall now describe some possible sources for heteroscedasticity as given in Gujarati et al. (2013).

(a) Error-learning models: As the number of hours put in typing practice increases, the average number of typing errors as well as their variance decreases.
(b) As income increases, not only savings increases but the variability in savings also increases—people have more choices with their income than to just save!
(c) Error variance changes with values of X due to some secondary issue.

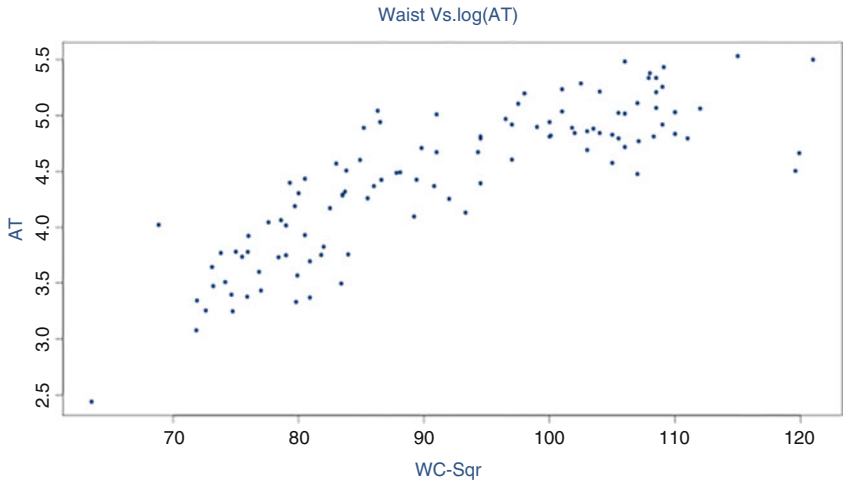


Fig. 7.17 Scatterplot of Waist vs Log AT

(d) Omitted variable: Due to the omission of some relevant regressor which is correlated with a regressor in the model, the omitted regressor remains in the error part and hence the error demonstrates a pattern when plotted against X—for example, in the demand function of a commodity, if you specify its own price but not the price of its substitutes and complement goods available in the market.
(e) Skewness in distribution: Distribution of income and education—bulk of income and wealth concentrated in the hands of a few.
(f) Incorrect data transformation, incorrect functional form specification (say linear X instead of Quadratic X, the true relation).

How does one detect heteroscedasticity? If you have a single regressor as in the case of Example 2.1 one can examine the scatterplot. If there are more regressors, one can plot the squared residuals against the fitted values and/or the regressors. If the plot is a random scatter, then you are fine with respect to the heteroscedasticity problem. Otherwise the pattern in the plot can give a clue regarding the nature of heteroscedasticity. For details regarding this and for formal tests for heteroscedasticity, we refer the reader to Gujarati et al. (2013).

Coming back to Example 2.1, one way to reduce the variation among adipose tissue values is by transforming AT to Log AT. Let us look at the scatterplot of Log AT vs Waist (Fig. 7.17).

We notice that the variation is more or less uniform across the range of Waist. However, we notice that there is a parabolic relationship between Log AT and Waist. So we fit a linear regression of Log AT on Waist and the square of Waist. The output is given below in Table 7.22.

The corresponding normal Q-Q plot is shown in Fig. 7.18.

Table 7.22 Regression output for the linear regression of Log AT on Waist and Square of Waist

```
> modellog<-lm(log(AT)~ Waist + Waist_sq, data=wc_at)
> summary(modellog)


Call:
lm(formula = log(AT) ~ Waist + Waist_sq, data = wc_at)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69843 -0.20915  0.01436  0.20993  0.90573

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.8240714  1.4729616  -5.312 6.03e-07 ***
Waist        0.2288644  0.0322008   7.107 1.43e-10 ***
Waist_sq    -0.0010163  0.0001731  -5.871 5.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.308 on 106 degrees of freedom
Multiple R-squared:  0.779,    Adjusted R-squared:  0.7748
F-statistic: 186.8 on 2 and 106 DF,  p-value: < 2.2e-16
```

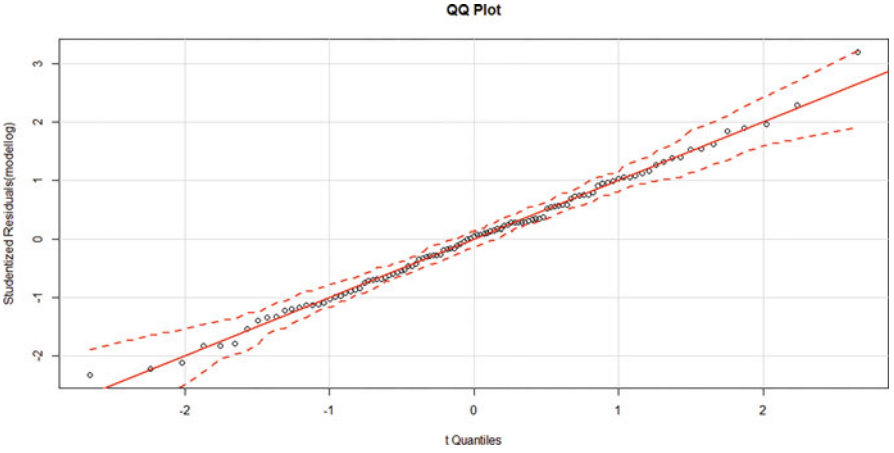


Fig. 7.18 Normal Q-Q plot of the standardized residuals in the linear regression of Log AT on Waist and Square of Waist

The corresponding plot of fitted values against residuals is given in Fig. 7.19.

Thus, both the plots tell us that the model is reasonable.

Alternatively, one can look at the linear regression of AT on Waist and look at the plot of squared residuals on Waist which is given in Fig. 7.20.

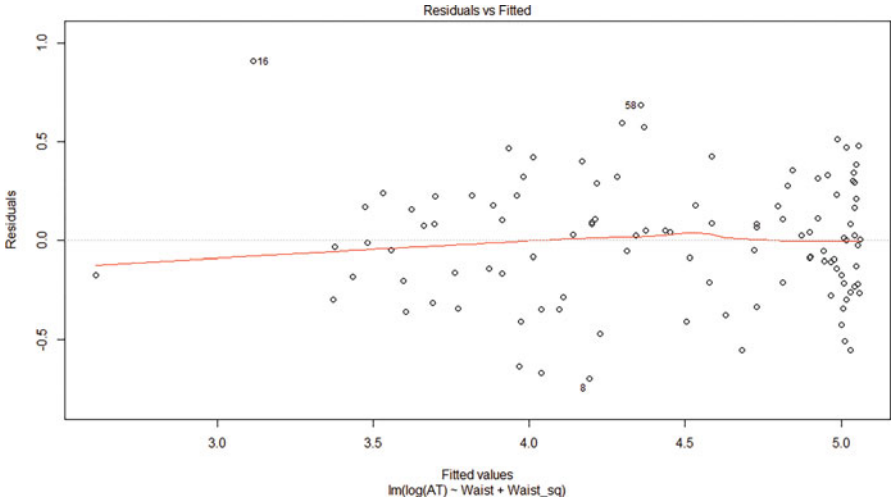
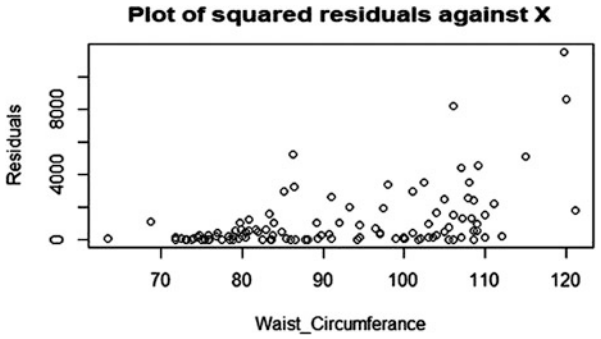


Fig. 7.19 Plot of fitted values vs residuals in the linear regression of log AT on Waist and Square of Waist

Fig. 7.20 Plot of Waist vs squared residuals in the regression of AT vs Waist



This shows a quadratic relationship and hence it suggests a regression of AT/Waist on 1/Waist (see Gujarati et al. 2013). We give the output of this regression in Table 7.23.

However, it can be checked that the corresponding normal Q-Q plot is not satisfactory.

Even though the usual least squares method, often called the ordinary least squares (OLS), is not appropriate, sometimes people use OLS with robust standard errors adjusted for heteroscedasticity. People also perform generalized least squares which can be performed as follows.

First run OLS. Get the residuals, $e_1, \ldots, e_N$. Divide the i^{th} row of the data by $\frac{|e_i|}{1-h_{ii}}$. Now run OLS on the new data. For details, refer to Gujarati et al. (2013).

Table 7.23 Regression output of AT/Waist on 1/Waist

```
> modelinv<-lm(ATdWaist ~ Waistin ,data=wc_at)
> summary(modelinv)

Call:
lm(formula = ATdWaist ~ Waistin, data = wc_at)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91355 -0.21062 -0.02604  0.17590  0.84371

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.5280     0.2158   16.35   <2e-16 ***
Waistin     -222.2786    19.1967  -11.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3305 on 107 degrees of freedom
Multiple R-squared:  0.5562,   Adjusted R-squared:  0.552
F-statistic: 134.1 on 1 and 107 DF,  p-value: < 2.2e-16
```

3.17 Model Validation

The linear regression model is developed using the sample on hand. Using the residual sum of squares, R^2, and other measures that we discussed, we can examine the performance of the model on the sample. But how does it perform on the population in general? How can we get an idea regarding this? For this purpose, if we have a reasonable size data, say of about 100 observations or more, about 20–30% of randomly selected observations from the sample are kept aside. These data are called *validation data*. The remaining data are called *training data*. The model is developed using the training data. The developed model's performance on the validation data is then checked using one or more of the following measures:

(a) Root Mean Square Error (RMSE)

We have the residual sum of squares, R_0^2 for the training data set. Using the developed model, predict the response variable for each of the observations in the validation data and compute the residual. Look at the residual sum of squares, $R_0^2(V)$ for the validation data set. Let N_1 and N_2 denote the number of observations in the training and validation data sets, respectively. The residual sum of squares per observation is called the RMSE. Let RMSE(T) and RMSE(V) denote the RMSE for training and validation data sets. Compare the RMSE for the training and validation data sets by computing

|RMSE (V) – RMSE(T)|/RMSE(T). If this is large, then the model does not fit well to the population. A thumb rule is to use a cutoff of 10%.

(b) Comparison of R^2

Define achieved R^2 for the validation data set as $R^2(V) = 1 - \frac{R_0^2(V)}{s^2}$, where s^2 is the sum of squared deviations of the response variable part of the validation data from their mean. Compare the achieved R^2 with R^2 of the training data set in the same manner as the RMSE above.

(c) Compare the box plots of the residuals in the training and validation data sets.

If we have a small data set, we can use the cross-validation method described as follows. Keep one observation aside and predict the response variable of this observation based on the model developed from the remaining observations. Get the residual. Repeat this process for all the observations. Let the sum of squared residuals be denoted by $R_0^2(C)$. Compute the cross-validation R^2 as

$$R^2(C) = 1 - \frac{R_0^2(C)}{\sum (y_i - \overline{y})^2}.$$

Interpret this the same way as R^2.

3.18 Broad Guidelines for Performing Linear Regression Modeling

There is no clear-cut algorithm for linear regression modeling. It is a combination of science, art, and technology. We give below broad guidelines one may follow in the pursuit of linear regression modeling for cross-sectional data (data collected on several units at one time point). We assume that we are dealing with data of reasonable size (as specified in Sect. 3.17).

Step 1. Keep aside a randomly chosen subset of about 20–30% observations. This will form the validation data. The remaining subset will be the training data. We shall develop the model based on the training data.

In steps 2–12, we work with the training data.

Step 2. Obtain the summary statistics of all the variables. This will help in understanding the measures of central tendency and dispersion of the individual variables.

Step 3. Obtain the box plots of all the variables. These will help in understanding the symmetry and skewness. Even if there is skewness, do not try to transform the variables to achieve symmetry at this stage. Remember that we need normality of the response variable conditional on the regressors and not necessarily the unconditional normality of the response variable. These plots may become helpful in making transformations at a later stage (see step 7).

Step 4. Obtain scatterplot matrix and correlation matrix of all the variables. We understand that the scatterplots and the correlations give us an idea of the linear relationship between two variables ignoring the impact of the other variables.

However, in the case of several regressors, we seek the impact of a regressor on the response variable after taking away the impact of other regressors on both the response variable and this regressor. A scatterplot matrix helps in understanding if there is collinearity between two regressors. It may give some broad idea of the relationship between pairs of the variables.

Step 5. Check whether there are categorical variables among the regressors. If so, following the guidelines given in Sect. 3.14, transform the categorical variables into appropriate dummy variables.

Step 6. Run a linear regression of the response variable on all the regressors. Check whether R square and adjusted R square are quite apart. If so, you know that there is issue with this model. Probably some unnecessary regressors are in the model or some observations are influential. These may become apparent in the later steps. Check whether some important regressors are insignificant or their coefficient estimates are of a wrong sign. If so, there may be a collinearity issue which will become clear when collinearity is examined in step 9.

Step 7. Obtain the residual plots—Fitted Values vs Residuals and Regressors vs Residuals. Follow the instructions in Sect. 3.10 for the necessary action. There can be occasions when the plots are not conclusive regarding the transformations. In such a case, one can look at the box plots for guidance. Sometimes these plots may indicate that some important variable correlated with the present set of regressors is missing. If so try to get data on a candidate variable based on the domain knowledge and follow the instructions in Sect. 3.11 to examine its usefulness and the form in which it should be included, if found suitable.

Step 8. Check for influential observations. Follow the instructions in Sect. 3.12.

Step 9. Check for collinearity among regressors. VIFs and variance decomposition proportions will help in detecting the collinear relationships. Follow the instructions in Sect. 3.13 for identifying collinear relationships and for remedial measures.

Step 10. Check for heteroscedasticity of the errors and take remedial actions as suggested in Sect. 3.16.

Step 12. Check for normality of the residuals and make a transformation on the response variable, if necessary, following the instructions in Sect. 3.15.

Step 13. Perform the validation analysis as in Sect. 3.17.

Step 14. If the model validation is successful, then fit a final regression following steps 6–12. Interpret the regression summary and translate your technological solution to the business problem into a business solution and prepare a report accordingly. If your model validation is not successful, then you are back to modeling and try fitting a suitable alternative model following steps 7–12.

One might wonder: where is the art in all this? Everything seems sequential. Often there is no unique transformation that the residual plots suggest. Again, there is no unique way of taking care of collinearity. Also, there is no unique way of tackling heteroscedasticity or nonnormality. In all such cases, several alternative models are indicated. The data scientist has to take a call based on his or her experience. Further, when collinearity is taken care of, one may find a new observation is influential. The reverse also is possible. So some iterations may be required.

4 FAQs

1. I have fitted a linear regression model to my data and found that $R^2 = 0.07$. Should I abandon performing regression analysis on this data set?

   Answer: There are several points to consider.

   (a) If R^2 is significantly different from 0 (based on the F-test), and if the assumptions are not violated, then one can use the model to study the impact of a significant regressor on the response variable when other regressors are kept constant.
   (b) If R^2 is not significantly different from 0, then this model is not useful. This is not to say that you should abandon the data set.
   (c) If the object is to predict the response variable based on the regressor knowledge of a new observation, this model performs poorly.
   (d) Suitable transformations may improve the model performance including R^2. (See the Exercise 7.1.) The basic principle is to look for simple models, as more complicated models tend to over-fit to the data on hand.

2. I have data for a sample on a response variable and one regressor. Why should I bother about regression which may not pass through any point when I can do a polynomial interpolation which passes through all the points leading to a perfect fit to the data?

   Answer: The data is related to a sample and our object is to develop a model for the population. While it is true that the polynomial interpolation formula is an exact fit to the data on hand, the formulae will be quite different if we bring in another observation or drop an observation. Thus, the model that we develop is not stable and is thus not suitable for the problem on hand. Moreover, the regressor considered is unlikely to be the only variable which impacts the response variable. Thus, there is error intrinsically present in the model. Interpolation ignores this. This is one of the reasons why over-fitting leads to problems.

3. I have performed the linear regression and found two of the regressors are insignificant. Can I drop them?

   Answer: There can be several reasons for a regressor to be insignificant:

   (a) This regressor may be involved in a collinear relationship. If some other regressor, which is also insignificant, is involved in this linear relationship, you may find the regressor insignificant. (See Table 7.9 where both *VOL* and *WT* are insignificant. Drop the variable *WT* and you will find that *VOL* is significant.)

(b) A variable in the current regression may not be significant. It is possible that if you bring in a new regressor, then in the presence of the new regressor, this variable may become significant. This is due to the fact that the correlation coefficient between a regressor and response variable can be smaller than the partial regression coefficient between the same two variables fixing another variable. (See Models 3 and 4 in Exercise 7.4.)
(c) Sometimes some unusual observations may make a regressor insignificant. If there are reasons to omit such observations based on domain knowledge, you may find the regressor significant.
(d) Sometimes moderately insignificant regressors are retained in the model for parsimony.
(e) Dropping a regressor may be contemplated as a last resort after exhausting all possibilities mentioned above.

4. I ran a regression of sales on an advertisement and some other regressors and found the advertisement's effect to be insignificant and this is not intuitive. What should I do?

   Answer: Examine points (a)–(c) in (3) above. Sometimes model misspecification can also lead to such a problem. After exhausting all these possibilities, if you still find the problem then you should examine whether your advertisement is highly uninspiring.

5. I have performed a linear regression. The residual plots suggested that I need to bring in the square term of a regressor also. I did that and once I included the square term and found from the variance decompositions proportions table that there is a collinearity among the intercept, the regressor, and its square. What should I do?

   Answer: Subtract an interpretable value close to the mean from the regressor and repeat the same with its square term. Usually this problem gets solved.

6. I have performed a linear regression and got the following estimated equation: $\widehat{Y} = -4.32 + 542.7X_1 - 3.84X_2 + 0.043X_3$. Can I conclude that the relative importance of X_1,X_2,X_3 is in that order?

   Answer: The value of the regression coefficient estimate depends on the scale of measurement of that regressor. It is possible that X_1,X_2,X_3 are measured in centimeters, meters, and kilometers, respectively. The way to assess is by looking at their t-values. There is also some literature on relative importance of regressors. One may see Kruskal (1987) and Gromping (2006).

7. How many observations are needed to perform a meaningful linear regression?

   Answer: The thumb rule is at least ten observations per estimated parameter. At least 30 observations are needed to estimate a linear regression model with two regressors and an intercept. Otherwise you may get significance by chance.

8. I made two separate regressions for studying the returns to education–one for men and the other for women. Can I compare the regression coefficients of both the models for education?

Answer: If you want to test the equality of coefficients in the two models, you need an estimate of the covariance between the two coefficient estimates, which cannot be obtained from separate regressions.

9. How do I interpret the intercept in a linear regression equation?
   Answer: We shall answer through a couple of examples.
   Consider the wage equation $\widehat{wage} = 14.3 + 2.83^{*}edu$, where *edu* stands for years of education. If you also have illiterates in the data used to develop the model, then the intercept 14.3 is the average wage of an illiterate (obtained from the wage equation by taking $edu = 0$). However, if your data is on individuals who have at least 7 years of education, do not interpret the intercept.
   Consider again the wage equation in Table 7.21. Here the intercept is the average wage of males having education of at most 12 years.

Electronic Supplementary Material

All the datasets, code, and other material referred in this section are available in www.allaboutanalytics.net.

- Data 7.1: AnscombesQuarter.csv
- Data 7.2: cars.csv
- Code 7.1: cars.R
- Data 7.3: cigarette_consumption.csv
- Code 7.2: cigarette_consumption.R
- Data 7.4: female.csv
- Code 7.3: female.R
- Data 7.5: healthcare1997.csv
- Data 7.6: leaf.csv
- Data 7.7: US_Dept_of_Commerce.csv
- Data 7.8: wage.csv
- Data 7.9: wc-at.csv
- Code 7.4: wc-at.R

Exercises

Ex. 7.1 Let the probability distribution of X be as follows:

Value	-2	-1	0	1	2
Probability	0.2	0.2	0.2	0.2	0.2

Define $Y = 3X^2+2$ and $Z = X^2$. Show that X and Y are uncorrelated. Show also that the correlation coefficient between Y and $Z = X^2$ is 1.

Ex. 7.2 Consider the following data on the variables X and Y.

Y	0.6	0.2	0.2	0.2	0.1	0.1	0.1	0.05	0.05	0.05
X	2.01	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0

(a) Do you use this X to predict Y?
(b) Can you guess the value of the correlation coefficient between X and Y?
(c) Now compute the correlation coefficient. Are you surprised?
(d) What happens to the correlation coefficient if 0.6 for Y is changed to 2.0 keeping the rest of the data unchanged? What can you conclude from this?
(e) What happens to the correlation coefficient if you change the value 2.01 for X to 50.0? How do you justify the answer?

Ex. 7.3 Consider data set 3 in Anscombe's quartet. Show that the observation 3 and 6 are influential but only 3rd observation is an outlier in the model $Y = \beta_0 + \beta_1 X + error$.

Ex. 7.4 Based on data on 100 adult males and 100 adult females on wage, years of education (YOE), age and sex (1 for male and 0 for female), perform the following linear regressions:

1. Model 1: Wage on age
2. Model 2: Wage on age and sex
3. Model 3: Wage on age and years of education
4. Model 4: Wage on age, sex, and years of education

In each case obtain also the residual plots including the normal Q-Q plot. Based on the analysis, answer the questions (a)–(j) given below.

(a) Notice that in models 1 and 3, age is insignificant. But it is significant in models 2 and 4. What is the reason for this?
(b) Interpret the residual plots: Normal Q-Q plot, age vs residual and fitted values vs residual in model 1.
(c) Based on your examination of the plots mentioned in (b), what action would you take and why?
(d) Compare the residual plots of models 1 and 3. What differences do you notice? How do you explain these differences?
(e) Consider the residual plots in model 4. You notice two clusters in the plot of fitted values vs residuals. However, there are no clusters in the residual plots of age vs residuals and YOE vs residuals. Is it strange? Give reasons for your answer.
(f) Does the output of model 3 indicate that there is collinearity between age and YOE? Give reasons for your answer.
(g) Compare the residual plots of YOE vs residuals in models 3 and 4. What do you think is the real reason for the difference? Do you believe that adding a square term of YOE in model 3 will improve the fit?

(h) In model 4, consider a male and a female of same age, 28, and the same education, of 10 years. Who earns more and by how much? Does it change if the age is 40 and the education is 5 years?
(i) If one wishes to test whether the females catch up with males with increasing level of education, how would you modify model 4 and what test do you perform?
(j) Notice that the Q-Q plots in models 1 and 3 are similar. Are they satisfactory? What about those in models 2 and 4? What is the reason for this dramatic change?

Ex. 7.5 In order to study the impact of bank deregulation on income inequality, yearly data was collected on the following for two states, say 1 and 0 during the years 1976 to 2006. Bank deregulation was enacted in state 1 and not in state 0. Gini index is used to measure income inequality. To control for time-varying changes in a state's economy, we use the US Department of Commerce data ("US_Dept_of_Commerce.csv") to calculate the growth rate of per capita Gross State Product (GSP). We also control for the unemployment rate, obtained from the Bureau of Labor Statistics, and a number of state-specific, time-varying socio-demographic characteristics, including the percentage of high school dropouts, the proportion of blacks, and the proportion of female-headed households.

Name	Description
Log_gini	Logarithm of Gini index of income inequality
Gsp_pc_growth	Growth rate of per capita Gross State Product (2000 dollars)
Prop_blacks	Proportion blacks
Prop_dropouts	Proportion of dropouts
Prop_female_headed	Proportion female-headed households
Unemployment	Unemployment
Post	Bank deregulation dummy
Treatment	Denoting two different states 1 and 0
Interaction	Post*treatment
Wrkyr	Year of study

Perform the linear regression of Log_gini on the rest of variables mentioned in the above table. Report your findings. How do you interpret the regression coefficient of the interaction? Are there any assumptions you are making over and above the usual linear regression assumptions?

Ex. 7.6 From the given dataset on health care outcomes ("healthcare1997.csv") in 1997 for many countries, you are required to develop a relationship between composite health care attainment measure (response variable) and the potential drivers of health care attainment. Develop the model on a training data set and examine the validity on the validation data.

The variables along with their descriptions are as follows:

COMP = Composite measure of health care attainment
DALE = Disability adjusted life expectancy (other measure)
HEXP = Per capita health expenditure
HC3 = Educational attainment
OECD = Dummy variable for OECD country (30 countries)
GINI = Gini coefficient for income inequality
GEFF = World Bank measure of government effectiveness*
VOICE = World Bank measure of democratization of the political process*
TROPICS = Dummy variable for tropical location
POPDEN = Population density*
PUBTHE = Proportion of health expenditure paid by public authorities
GDPC = Normalized per capita GDP

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician, 27*, 17–21.
Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics*. New York: John Wiley and Sons.
Brundavani, V., Murthy, S. R., & Kurpad, A. V. (2006). Estimation of deep-abdominal-adipose-tissue (DAAT) accumulation from simple anthropometric measurements in Indian men and women. *European Journal of Clinical Nutrition, 60*, 658–666.
Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (5th ed.). New York: John Wiley and Sons.
Current Population Survey. (1994) United States Department of Commerce. Bureau of the Census.
Despres, J. P., Prud'homme, D., Pouliot, A. T., & Bouchard, C. (1991). Estimation of deep abdominal adipose-tissue accumulation from simple anthropometric measurements in men. *American Journal of Clinical Nutrition, 54*, 471–477.
Draper, N., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley and Sons.
Greene, W. H. (2012). *Econometric analysis* (7th ed.). London: Pearson Education.
Gromping, U. (2006). Relative importance for linear regression in R. *Journal of Statistical Software, 17*, 1–27.
Gujarati, D. N., Porter, D. C., & Gunasekar, S. (2013). *Basic econometrics* (5th ed.). New Delhi: Tata McGrawHill.
Kruskal, W. (1987). Relative importance by averaging over orderings. *American Statistician, 41*, 6–10.
Thode, H. C., Jr. (2002). *Testing for normality*. New York: Marcel Dekker.
Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *Journal of Royal Statistical Society, Series B, 58*, 267–288.