Chapter 13
Logistic Regression: Developing a Model for Risk Analysis

Learning Objectives

After completing this chapter, you should be able to do the following:

- Learn the difference between logistic regression and ordinary least squares regression.
- Know the situation where logistic regression can be used.
- Describe the logit transformation used in the analysis.
- Understand different terminologies used in logistic regression.
- Explain the steps involved in logistic regression.
- Understand the assumptions used in the analysis.
- Know the SPSS procedure involved in logistic regression.
- Understand the odds ratio and its use in interpreting the findings.
- Interpret the outputs of logistic regression generated by the SPSS.

Introduction

Logistic regression is a useful statistical technique for developing a prediction model for any event that is binary in nature. A binary event can either occur or not occur. It has only two states which may be represented by 1(occurrence) and 0 (nonoccurrence). Logistic regression can also be applied in a situation where the dependent variable has more than two classifications. The logistic regression can either be binary or multinomial depending upon whether the dependent variable is classified into two groups or more than two groups, respectively. In this chapter, the discussion shall be made only for binary logistic regression.

Logistic regression is useful in a situation where we are interested to predict the occurrence of any happening. It has vast application in the area of management, medical and social researches because in all these discipline occurrence of a phenomenon depends upon the independent variables that are metric as well as categorical in nature. Logistic regression can be used for developing model for

J.P. Verma, *Data Analysis in Management with SPSS Software*,
DOI 10.1007/978-81-322-0786-3_13, © Springer India 2013

financial prediction, bankruptcy prediction, buying behavior, fund performance, credit risk analysis, etc. We have witnessed the failure of high-profile companies in the recent past. This has generated an interest among the industrial researcher to develop a model for bankruptcy prediction. Such model can also be made for retail and other firms on the basis of the accounting variables such as inventories, liabilities, receivables, net income (loss), and revenue. On the basis of such model, one can estimate the risk of bankruptcy of any organization.

In hilly regions, there is always a fear of landslide which causes heavy damage to the infrastructure and human lives. The logistic regression model can be used to find out the landslide susceptibility in such areas. The model can identify more probable areas prone to landslides. On the basis of such information, appropriate measures may be taken to reduce the risk from potential landslide hazard. In developing the logistic model for landslide susceptibility, the remote sensing and geographic information system (GIS) data may be used as independent variables.

In product marketing, it is required to identify those customers on whom advertisement should be focused. Consider a situation in which a company has introduced an herbal cream costing Rs. 520 and wishes to identify the parameters responsible for the customers to buy this product. The data on the parameters like age, gender, income, and family size may be collected on the customers who have inspected the cream at the counter in few stores. Here the dependent variable is the buying decision of the customer (1 if the cream is purchased and 0 if not), whereas the independent variables are the mix of ratio (age, income, family size) and categorical variable (gender). Since the dependent variable is the dichotomous variable and independent variables have a combination of ratio and categorical variables, the logistic regression can be applied to identify the variables that are responsible for the buying behavior of the customers. Further, the relative importance of the independent variable can also be known by this analysis, and therefore, decision maker may focus on those variables which maximize the chances of buying the product.

In the financial sector, financial companies may be interested to find the attributes of the financial managers responsible for fund performance. One may investigate by using logistic model as to which of the independent variables out of educational background, gender, and seniority of the fund managers are related with the fund performance.

Due to large number of listed companies on the bourses, there is always a fear of credit issues and frequent credit crises. The logistic model may be developed for credit risk analysis which may provide the monitoring agency a system of identifying corporate financial risk which works as an effective indicator system. In developing such model, the past data is usually taken on the identified parameters.

What Is Logistic Regression?

Logistic regression is a kind of predictive model that can be used when the dependent variable is a categorical variable having two categories and independent variables are either numerical or categorical. Examples of categorical variables are

buying/not buying a product, disease/no disease, cured/not cured, survived/not survived, etc. The logistic regression is also known as logit model or logistic model. The dependent variable in the logit model is often termed as outcome or target variable, whereas independent variables are known as predictive variables. A logistic regression model is more akin to nonlinear regression such as fitting a polynomial to a set of data values. By using the logistic model, the probability of occurrence of an event is predicted by fitting data to a logit function or a logistic curve.

Important Terminologies in Logistic Regression

Before getting involved into serious discussion about the logistic regression, one must understand different terminologies involved in it. The terms which are required in understanding the logistic regression are discussed herewith.

Outcome Variable

Outcome variable is that variable in which a researcher is interested. In fact it is a dependent variable which is binary in nature. The researcher is interested to know the probability of its happening on the basis of several risk factors. For example, the variables like buying decision (buying = 1, not buying = 0), survival (surviving = 1, not surviving = 0), bankruptcy (bankruptcy of an organization = 1, no bankruptcy = 0), and examination results (pass =1, fail = 0) are all outcome variables.

Natural Logarithms and the Exponent Function

The natural log is the usual logarithmic function with base e. The natural log of X is written as $\log(X)$ or $\ln(X)$. On the other hand, the exponential function involves the constant "e" whose value is equal to 2.71828182845904 ($\approx$2.72). The exponential of X is written as $\exp(x) = e^x$. Thus, $\exp(4)$ equals to $2.72^4 = 54.74$.

Since natural log and exponential function are opposite to each other,

$$E^4 = 54.74$$

$$\Rightarrow \quad \ln(54.74) = 4$$

Odds Ratio

If probability of success (p) of any event is 0.8, then the probability of its failure is $(1 - p) = 1-0.8 = 0.2$. The odds of the success can be defined as the ratio of the probability of success to the probability of failure. Thus, in this example, odds of success is $0.8/0.2 = 4$. In other words, the odds of success is 4 to 1. If the probability of success is 0.5, then the odds of success is 1 and it may be concluded that the odds of success is 1 to 1.

In logistic regression, odds ratio can be obtained by finding the exponential of regression coefficient, exp(B), and is sometimes written as eB. If the regression coefficient B is equivalent to 0.80, then the odds ratio will be 2.40 because exp (0.8) = 2.4.

The odds ratio of 2.4 indicates that the probability of Y equals to 1 is 2.4 times as likely as the value of X is increased by one unit. If an odds ratio is .5, it indicates that the probability of $Y = 1$ is half as likely with an increase of X by one unit (here there is a negative relationship between X and Y). On the other hand, the odds ratio 1.0 indicates that there is no relationship between X and Y.

The odds ratio can be better understood if both variables Y and X are dichotomous. In that case, the odds ratio can be defined as the probability that Y is 1 when X is 1 compared to the probability that Y is 1 when X is 0. If the odds ratio is given, then B coefficient can be obtained by taking the log of the odds ratio. It is so because log and exponential functions are opposite to each other.

The transformation from probability to odds is a monotonic transformation. It means that the odds increases as the probability increases or vice versa. Probability ranges from 0 to 1, whereas the odds ranges from 0 to positive infinity.

Similarly the transformation from odds to log of odds, known as log transformation, is also a monotonic transformation. In other words, the greater the odds, the greater is the log of odds and vice versa. Thus, if the probability of success increases, the odds ratio and log odds both increase and vice versa.

Maximum Likelihood

Maximum likelihood is the method of finding the least possible deviation between the observed and predicted values using the concept of calculus specifically derivatives. It is different than ordinary least squares (OLS) regression where we simply try to find the best-fitting line by minimizing the squared residuals.

In maximum likelihood (ML) method, the computer uses different "iterations" where different solutions are tried for getting the smallest possible deviations or best fit. After finding the best solution, the computer provides the final value for the deviance, which is denoted as "−2 log likelihood" in SPSS. Cohen et al. (2003) called this deviance statistic as −2LL, whereas some other authors like Hosmer and Lemeshow (1989) called it D. This deviance statistic follows the chi-square distribution.

The likelihood ratio test, D, is used as goodness-of-fit. This test is referred in SPSS by "chi-square." The significance of this test can be seen by looking to its value in the chi-square table in the appendix using degrees of freedom equal to the number of predictors.

Logit

The logit is a function which is equal to the log odds of a variable. If p is a probability that $Y = 1$(occurrence of an event), then $p/(1 - p)$ is the corresponding odds. The logit of the probability p is given by

$$\text{Logit}(\text{p}) = \log\left(\frac{\text{p}}{1-\text{p}}\right) \tag{13.1}$$

In logistic regression, logit is a special case of a link function. In fact, this logit serves as a dependent variable and is estimated from the model.

Logistic Function

A logistic curve is just like sigmoid curve and is obtained by the logistic function given by

$$p = f(z) = \frac{\text{e}^Z}{1+\text{e}^Z} \tag{13.2}$$

The shape of the curve is like a letter "S." In logistic function, the argument z is marked along horizontal axis and the value of the function $f(z)$ along the vertical axis (Fig. 13.1).

The main feature of this logistic function is that the variable Z can assume any value from minus $-\infty$ to $+\infty$, but the outcome variable p can have the values only in the range 0–1. This function is used in logistic regression model to find the probability of occurring the target variable for a given value of independent variables.

Logistic Regression Equation

The logistic regression equation is similar to the ordinary least squares (OLS) regression equation with the only difference that the dependent variable here is

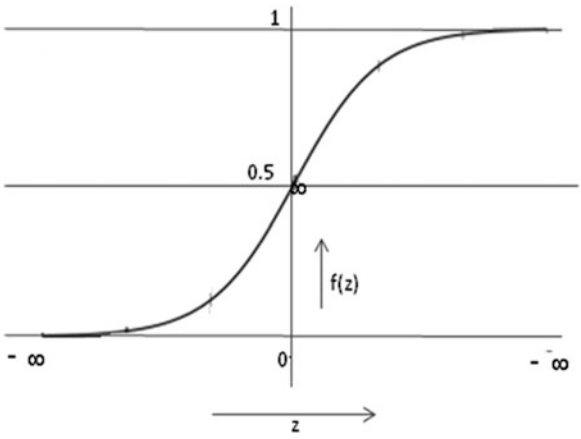


Fig. 13.1 Shape of the logistic function

the log odds of the probability that the dependent variable $Y = 1$. It is written as follows:

$$\text{logit} = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X_1 + B_2X_2 + \ldots\ldots\ldots + B_nX_n \tag{13.3}$$

where B_0 is an intercept and $B_1, B_2 \ldots B_n$ are the regression coefficients of $X_1, X_2 \ldots \ldots, X_n$, respectively. The dependent variable in logistic regression is log odds, which is also known as logit.

Since in logistic regression log odds acts as a dependent variable which is regressed on the basis of the independent variables, interpretation of regression coefficients is not as easy as in case of OLS regression. In case of OLS regression, the regression coefficient b represents the change in Y with one unit change in X. This concept is not valid in case of logistic regression equation; instead the regression coefficient b is converted into odds ratio to interpret the happening of outcome variable. The interpretation of odds ratio has been discussed above in detail under the heading "Odds Ratio."

Judging the Efficiency of the Logistic Model

In case of OLS regression equation, R^2 used to be the measure of efficiency in assessing the suitability of the model. But in case of logistic regression, this statistic is no longer valid indicator of model robustness, because of the fact that the dependent variable here is a binary variable. Thus, to assess the suitability of the logistic model, we use the concept of deviance. In logistic regression, the chi-square

is used as a measure of model fit instead of R^2. It tells you about the fit of the observed values (Y) to the expected values ($\hat{Y}$). If the difference between the observed values from the expected values increases, the fit of the model becomes poorer. Thus, the effort is to have the deviance as small as possible. If more relevant variables are added to the equation, the deviance becomes smaller, indicating an improvement in fit.

Understanding Logistic Regression

In logistic regression, the approach of prediction is similar to that of ordinary least squares (OLS). However, in logistic regression, a researcher predicts the probability of an occurrence of a dependent variable which is binary in nature. Another difference in logistic regression is that the independent variables can be a mix of numerical and categorical. Due to dichotomous nature of the dependent variable, assumptions of OLS that the error variances (residuals) are normally distributed are not satisfied. Instead, they are more likely to follow a logistic distribution.

In using logistic distribution, one needs to make an algebraic conversion to arrive at usual linear regression equation. In logistic regression, no standard solution is obtained and no straightforward interpretation can be made as is done in case of OLS regression. Further, in logistic model, there is no R^2 to measure the efficiency of the model; rather a chi-square test is used to test how well the logistic regression model fits the data.

Graphical Explanation of Logistic Model

Let us first understand the concept of logistic regression with one independent variable. Consider a situation where we try to predict whether a customer would buy a product(Y) depending upon the number of days(X) he saw the advertisement of that product. It is assumed that the customers who watch the advertisement for many days will be more likely to buy the product. The value of Y can be 1 if the product is purchased by the customer and 0 if not.

Since the dependent variable is not a continuous, hence the goal of logistic regression is to predict the likelihood that Y is equal to 1 (rather than 0) given certain values of X. Thus, if there is a positive linear relationship between X and Y, then the probability that a customer will buy the product ($Y = 1$) will increase with the increase in the value of X (number of days advertisement seen). Hence, we are actually predicting the probabilities instead of value of the dependent variable.

Table 13.1 Mean score for each category

No. of days advertisement viewing	Probability that $Y = 1$ (average of 0s and 1s in each category)
0–3	.17
4–6	.40
7–9	.50
10–12	.56
13–15	.96

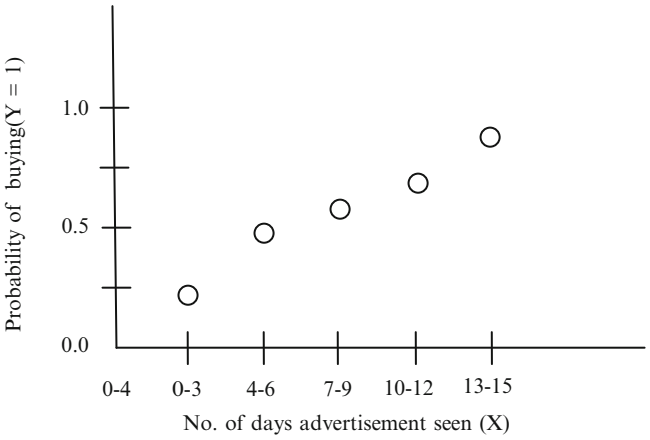


Fig. 13.2 Graphical representation of the probability of buying versus number of days advertisement seen

In this simulated experiment in investigating the behavior of 100 customers in terms of buying the product of more than Rs. 1,000, their range of viewing the advertisement for the number of days was from 0 to 15 days. We may plot the probability that $Y = 1$ with the increase in the value of X in terms of the graph. To make it more convenient, let us club the number of advertisement-viewing days into the categories 0–3, 4–6, 7–9, 10–12, and 13–15. Computing the mean score on Y (taking the average of 0s and 1s) for each category, the data would look like as shown in Table 13.1.

If we plot these data, the graph would look like as shown in Fig. 13.2.

If we look at this graph, it looks like an S-shaped graph. If there is a strong relationship between X and Y, the graph would be closer to perfect S-shaped unlike the OLS regression where you get the straight line.

Logistic Model with Mathematical Equation

If Y is the target variable (dependent) and X is the predictive variable and if the probability that $Y = 1$ is denoted as $\hat{p}$, then the probability that Y is 0 would be $1 - \hat{p}$. The logistic model for predicting $\hat{p}$ would be given by

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1 X \tag{13.4}$$

where $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$ is the log of the odds ratio and is known as logit and B_0 is the constant and B_1 is the regression coefficient.

In effect, in logistic regression this logit, $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$, is the dependent variable against which independent variables are regressed.

From Eq. (13.4), the probability ($\hat{p}$) that $Y = 1$ can be computed for a given value of X.
Let us assume that

$$Z = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1 X \tag{13.5}$$

$$\Rightarrow \qquad \frac{\hat{p}}{1-\hat{p}} = e^Z$$

Or

$$\hat{p} = \frac{e^Z}{1+e^Z} = \frac{e^{B_0+B_1X}}{1+e^{B_0+B_1X}} \tag{13.6}$$

Thus, in the logistic regression, first a logit or log of odds ratio, that is, $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$, is computed for a given value of X, and then the probability ($\hat{p}$) that $Y = 1$ is computed by using formula (13.6). In fact (13.6) gives the logistic function as

$$f(z) = \frac{e^Z}{1+e^Z} \tag{13.7}$$

This function if plotted by taking z on horizontal axis and $f(z)$ on vertical axis looks like as shown in Fig. 13.3.

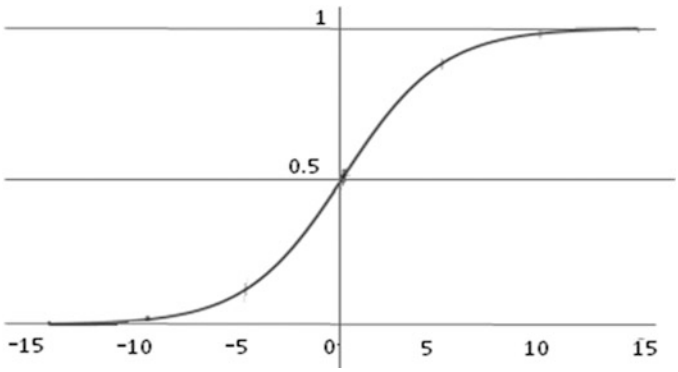


Fig. 13.3 Logistic function for finding the probability of $Y = 1$

Interpreting the Logistic Function

In logistic regression, the logistic function shown in Fig. 13.3 is used for estimating the probability of an event happening ($Y = 1$) for different values of X. Let us see how it is done.

In the logistic function shown in (13.7), the input is z and output is $f(z)$. The value of z is estimated by the logistic regression Eq. (13.5) on the basis of the value of X. The important characteristics of the logistic function are that it can take any value from negative infinity to positive infinitive, but the output will always be in the range of 0–1.

If there are n independent variables, then the value of z or logit or log of odds shall be estimated by the equation:

$$Z = \text{logit} = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X_1 + B_2X_2 + \ldots\ldots\ldots + B_nX_n \qquad (13.8)$$

where B_0 is an intercept and B_1, B_2. . ..B_n are the regression coefficients of X_1, X_2 . . ...,X_n, respectively.

The variable Z is estimated from (13.8) for a given value of Xs. It is a measure of the total contribution of all the independent variables used in the model.

If the outcome variable is the risk factor for happening of an event say bankruptcy of an organization, then each of the regression coefficients shows the contribution toward the probability of that outcome. If the regression coefficient is positive, it indicates that the explanatory variable increases the probability

of the outcome, whereas in case of negative regression coefficient, it decreases the probability of that outcome. On the other hand, a large regression coefficient means that the corresponding variable is a high risk factor which strongly influences the probability of that outcome, whereas a near-zero regression coefficient indicates that the corresponding variable is not an important risk factor and has little influence on the probability of that outcome.

Assumptions in Logistic Regression

Following are the assumptions used in the logistic regression:

1. The target variable is always binary. If by nature it is continuous, a criterion may be defined to convert it into binary.
2. The predictor variables can either be numerical or categorical. In case the categorical variable has more than two categories, a dummy variable D (it may have the variable name as well) is created and different categories may be denoted by code 1, 2, 3, etc. Care should be taken that the highest code should refer the reference category. By default, the SPSS assumes the highest coding as reference category and marks it 0. For instance, if the qualification is taken as categorical variable, then this variable D may be coded as follows:

   $D = 3$, if the subject's qualification is XII standard or less.
   $D = 2$, if the subject is graduate.
   $D = 1$, if the subject's qualification is postgraduation or more.

   In SPSS the highest coding is taken as the reference category by default, and therefore, you will find that in the output, XII or less qualification category is represented by 0 and the interpretation is made with reference to this category only. However, SPSS does provide the facility to change the reference category to the lowest code as well.
3. It is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables.
4. Many authors suggested that a minimum of ten events per predictive variables should be taken in the logistic regression. For example, in a study where cure is the target variable of interest and 100 out of 150 patients get cured, the maximum number of independent variables one can have in the model is $100/10 = 10$.

Important Features of Logistic Regression

1. The logistic regression technique is more robust because the independent variables do not have to be normally distributed or have equal variance in each group.

2. The independent variables are not required to be linearly related with the dependent variable.
3. It can be used with the data having nonlinear relationship.
4. The dependent variable need not follow normal distribution.
5. The assumption of homoscedasticity is not required. In other words, no homogeneity of variance assumption is required.

Although the logistic regression is very flexible and can be used in many situations without imposing so many restrictions on the data set, the advantages of logistic regression come at a cost. It requires large data set to achieve reliable and meaningful results. Whereas in OLS regression and discriminant analysis, 5 to 10 data per independent variable is considered to be minimum threshold, logistic regression requires at least 50 data per independent variable to achieve the reliable findings.

Research Situations for Logistic Regression

Due to the flexibility about its various assumptions, the logistic regression is widely used in many applications. Some of the specific applications are discussed below:

1. A food joint chain may be interested to know as to what factors may influence the customers to buy big-size Pepsi in the fast-food center. The factors may include the type of pizza (veg. or non-veg.) ordered, whether French fries ordered, the age of the customer, and their body size (bulky or normal). The logistic model can provide the solution in identifying the most probable parameters responsible for buying big-size Pepsi in different food chains.
2. A study may investigate the parameters responsible for getting admission to MBA program in Harvard Business School. The target variable is a dichotomous variable with 1 indicating the success in getting admission, whereas 0 indicates failure. The parameters of interest may be working experience of the candidates in years, grades in the qualifying examination, TOEFL and GMAT scores, and scores on the testimonials. By way of logistic model, the relative importance of the independent variables may be identified and the probability of success of an individual may be estimated on the basis of the known values of the independent variables.
3. A market research company may be interested to investigate the variables responsible for a customer to buy a particular life insurance cover. The target variable may be 1 if the customer buys the policy and 0 if not. The possible independent variables in the study may be the age, gender, socioeconomic status, family size, profession (service/business), etc. By knowing the most likely causes for getting success in selling the policy, the company may target the campaign toward the target audience.

4. Incidence of HIV infection may be investigated by using the logistic model, where the independent variables may be identified as person's movement (frequent or less frequent), age, sex, occupation, personality type, etc. The strategy may be developed by knowing the most dominant causes responsible for HIV infection, and accordingly mass campaign may be initiated for different sections of the society in an efficient manner. One of the interesting facts in such studies may be to investigate the important factors of the HIV incidences in different sections of the society due to different dynamics.
5. The incidence of cardiac death may be investigated based on the factors like age, sex, activity level, BMI, and blood cholesterol level of the patients by fitting the logit model. The odds ratio will help you find the relative magnitude of risk involved with different factors.

Steps in Logistic Regression

After understanding the concepts involved in logistic regression, now you are ready to use this analysis for your problem. The detailed procedure of this analysis using SPSS shall be discussed by using a practical example. But before that, let us summarize the steps involved in using the logistic regression:

1. Define the target variable and code it 1 if the event occurs and 0 otherwise. The target variable should always be dichotomous.
2. Identify the relevant independent variables responsible for the occurrence of target variable.
3. In case if any independent variable is categorical having more than two categories, define the coding for different categories as discussed in the "Assumptions" section.
4. Develop a regression model by taking dependent variable as log odds of the probability that target variable $Y = 1$. Logistic regression model can be developed either by using forward/backward step methods or by using all the independent variables in the model. Forward/backward step methods are usually used in explorative study where it is not known whether the independent variable has some effect on the target variable or not. On the other hand, all the independent variables are used in developing a model if the effect of independent variables is known in advance and one tries to authenticate the model. Several options for forward/backward methods are available in the SPSS, but "Forward:LR" method is considered to be the most efficient method. On the other hand, for taking all the independent variables in the model, the SPSS provides a default option with "Enter" command.
5. After choosing the method for binary logic regression, the model would look like as follows where $\hat{p}$ is the probability that the target variable $Y = 1$:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = Z = B_0 + B_1X_1 + B_2X_2 + \ldots\ldots\ldots + B_nX_n$$

The variables have their usual meanings. The log odds $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$ is also known as logit.

6. The estimated probability of occurring the target variable can be estimated for a given set of values of independent variables by using the following formula:

$$\widehat{p} = \frac{e^Z}{1+e^Z} = \frac{e^{B_0+B_1X_1+B_2X_2\ldots..B_nX_n}}{1+e^{B_0+B_1X_1+B_2X_2+\ldots...B_nX_n}}$$

The above-mentioned equation gives rise to the logistic curve which is S-shaped as shown in Fig. 13.3. The probability can also be computed from this curve by computing the value of Z.

7. Exponential of the regression coefficient is known as odds ratio. These odds ratios are used to find the relative contribution of all the independent variables toward the occurrence of target variable. Thus, the odds ratio corresponding to each of the regression coefficients is computed for investigating the relative contribution of independent variables toward the occurrence of dependent variable. For example, the odds ratio of 3.2 for the variable X_1 indicates that the probability of Y (dependent variable) equals to 1 is 3.2 times as likely as the value of X_1 is increased by one unit. And if an odds ratio for the variable X_3 is .5, it indicates that the probability of $Y = 1$ is half as likely with an increase of X_3 by one unit (here there is a negative relationship between X_3 and Y). On the other hand, if the odds ratio for the variable X_2 is 1.0, it indicates that there is no relationship between X_2 and Y.

Solved Example of Logistics Analysis Using SPSS

Example 13.1 A researcher wanted to investigate the factors responsible for getting the job of coin note examiner in banks. The data was obtained by the recruitment agency that was responsible for appointment of bank employees. The investigator collected the data on the outcome variable (appointed or not appointed) and independent variables like education (number of years of college education), sex, experience, age, metro/nonmetro status, and marital status. These data are shown in Table 13.2. Apply logistic regression by using SPSS to develop a model for estimating the probability of success in getting the job on the basis of candidate's profiles. Further, discuss the comparative importance of these independent variables in getting success during an interview for the job. The coding for the categorical variables is shown below the table.

Table 13.2 Data on the candidate's profile along with success status

S.N.	Job success	Education	Sex	Experience in years	Age	Metro	Marital status
1	1	16	1	7	23	1	1
2	0	15	1	5	25	0	0
3	1	16	1	5	27	1	1
4	1	15	1	2	26	1	0
5	0	16	0	3	28	0	0
6	1	15	1	2	26	0	1
7	0	13	1	3	33	1	1
8	0	12	0	2	32	0	1
9	1	12	1	3	26	1	1
10	0	13	0	3	30	0	0
11	1	12	0	1	28	1	1
12	0	12	0	2	28	0	0
13	1	15	1	6	32	1	1
14	1	12	1	3	38	0	1
15	0	16	0	2	23	0	0
16	1	15	1	3	22	1	0
17	1	16	1	7	23	0	1
18	0	15	1	5	25	0	0
19	1	16	1	5	27	1	1
20	1	12	0	2	28	0	1
21	1	16	1	4	28	1	0
22	1	15	1	3	28	0	1
23	0	12	0	2	26	1	0
24	0	14	0	5	29	0	0

Job success : 0 : Failure 1 : Success
Sex : 0 : Female 1 : Male
Metro : 0 : Nonmetro resident 1 : Metro resident
Marital status : 0 : Unmarried 1 : Married

Solution

The above-mentioned problem can be solved by using SPSS. The steps involved in getting the outputs shall be discussed first and then the output so generated shall be explained to fulfill the objectives of the study.

The logistic regression in SPSS is run in two steps. The outputs generated in these two sections have been discussed in the following two steps:

First Step

Block 0: Beginning Block

The first step, called Block 0, includes no predictors and just the intercept. This model is developed by using only constant and no predictors. The logistic regression compares this model with a model having all the predictors to assess whether the later model is more efficient. Often researchers are not interested in this model. In this part, a "null model," having no predictors and just the intercept, is described.

Because of this, all the variables entered into the model will figure in the table titled "Variables not in the Equation."

Second Step

Block 1: Method = Forward:LR

The second step, called Block 1, includes the information about the variables that are included and excluded from the analysis, the coding of the dependent variable, and coding of any categorical variables listed on the categorical subcommand. This section is the most interesting part of the output in which generated outputs are used to test the significance of the overall model, regression coefficients, and odds ratios.

The above-mentioned outputs in two steps are generated by the SPSS through a single sequence of commands, but the outputs are generated in two different sections with the headings "Block 0: Beginning Block" and "Block 1." You have the liberty to use any method of entering independent variables in the model out of different methods available in SPSS. These will be discussed while explaining screen shots of logistic regression in the next section.

The procedure of logistic regression in SPSS shall be defined first and then relevant outputs shall be shown with explanation.

SPSS Commands for the Logistic Regression

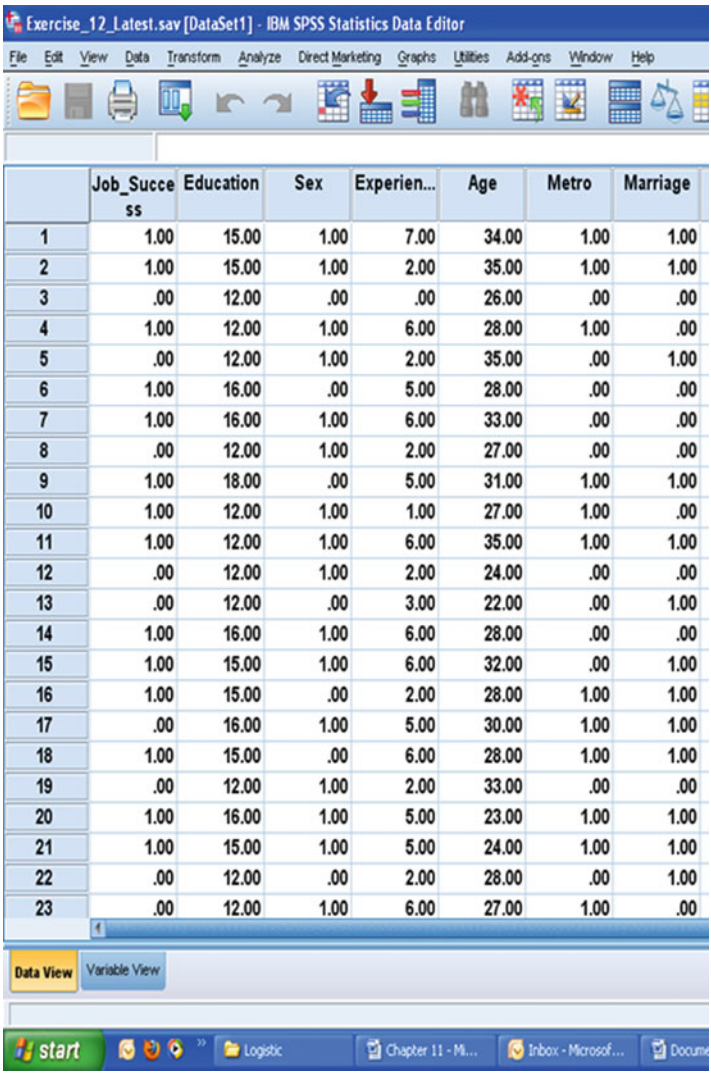
To run the commands for logistic regression, a data file is required to be prepared. The procedure for preparing the data file has been explained in Chap. 1. After preparing the data file, do the following steps for generating outputs in logistic regression:

(i) *Data file:* In this problem, job success is a dependent variable which is binary in nature. Out of six independent variables, three variables, namely, sex, metro, and marital status, are binary, whereas remaining three, education, experience, and age, are scale variables. In SPSS all binary variables are defined as nominal. After preparing the data file by defining variable names and their labels, it will look like as shown in Fig. 13.4.

(ii) *Initiating command for logistic regression:* After preparing the data file, click the following commands in sequence (Fig. 13.5):

Analyze → Regression → Binary˜Logistic

(iii) *Selecting variables for analysis:* After clicking the **Binary Logistic** option, you will get the next screen for selecting dependent and independent variables. After selecting all the independent variables, you need to select the binary independent variables included in it by clicking the option. The selection of variables can be made by following the below-mentioned steps:



	Job_Success	Education	Sex	Experien...	Age	Metro	Marriage
1	1.00	15.00	1.00	7.00	34.00	1.00	1.00
2	1.00	15.00	1.00	2.00	35.00	1.00	1.00
3	.00	12.00	.00	.00	26.00	.00	.00
4	1.00	12.00	1.00	6.00	28.00	1.00	.00
5	.00	12.00	1.00	2.00	35.00	.00	1.00
6	1.00	16.00	.00	5.00	28.00	.00	.00
7	1.00	16.00	1.00	6.00	33.00	.00	.00
8	.00	12.00	1.00	2.00	27.00	.00	.00
9	1.00	18.00	.00	5.00	31.00	1.00	1.00
10	1.00	12.00	1.00	1.00	27.00	1.00	.00
11	1.00	12.00	1.00	6.00	35.00	1.00	1.00
12	.00	12.00	1.00	2.00	24.00	.00	.00
13	.00	12.00	.00	3.00	22.00	.00	1.00
14	1.00	16.00	1.00	6.00	28.00	.00	.00
15	1.00	15.00	1.00	6.00	32.00	.00	1.00
16	1.00	15.00	.00	2.00	28.00	1.00	1.00
17	.00	16.00	1.00	5.00	30.00	1.00	1.00
18	1.00	15.00	.00	6.00	28.00	1.00	1.00
19	.00	12.00	1.00	2.00	33.00	.00	.00
20	1.00	16.00	1.00	5.00	23.00	1.00	1.00
21	1.00	15.00	1.00	5.00	24.00	1.00	1.00
22	.00	12.00	.00	2.00	28.00	.00	1.00
23	.00	12.00	1.00	6.00	27.00	1.00	.00

Fig. 13.4 Screen showing data file for the logistic regression analysis in SPSS

- Select the dependent variable from the left panel to the "Dependent" section in the right panel.
- Select all independent variables including categorical variables from left panel to the "Covariates" section in the right panel.
- Click the command **Categorical** and select the categorical variables from the "Covariates" section to the "Categorical Covariates" in the right panel. The screen will look like Fig. 13.6.
- Click ***Continue.***

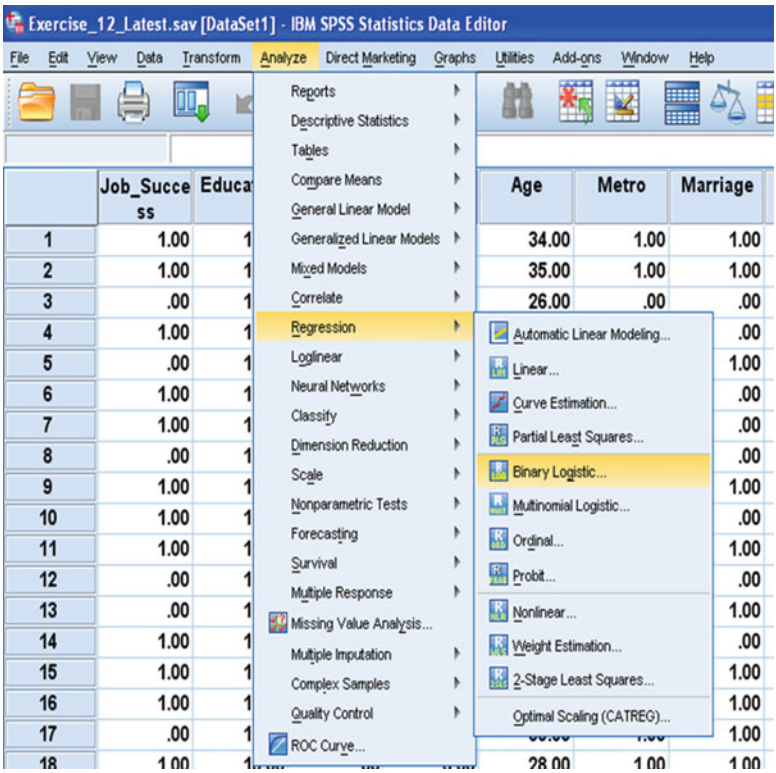


Fig. 13.5 Screen showing of SPSS commands for logistic regression

(iv) *Selecting options for computation:* After selecting the variables, you need to define different options for generating the outputs in logistic regression. Do the following steps:

- Click the tag **Options** in the screen shown in Fig. 13.6 for selecting outputs related to statistics and plots. Do the following steps:
- Check "Classification Plots."
- Check "Hosmer-Lemeshow goodness-of-fit."
- Let all other default options be selected. The screen will look like Fig. 13.7.
- Click ***Continue***.

(v) *Selecting method for entering independent variables in logistic regression*: You need to define the method of entering the independent variables for developing the model. You can choose any of the options like Enter, Forward:LR, Forward:Wald, Backward:LR, or Backward:Wald. Enter method is usually selected when a specific model needs to be tested or the contribution of independent variables toward the target variable is known in advance. On the other hand, if the study is exploratory in nature, then any of the forward or backward methods can be used. In this study, the Forward:LR method shall be used because the study is exploratory in nature.

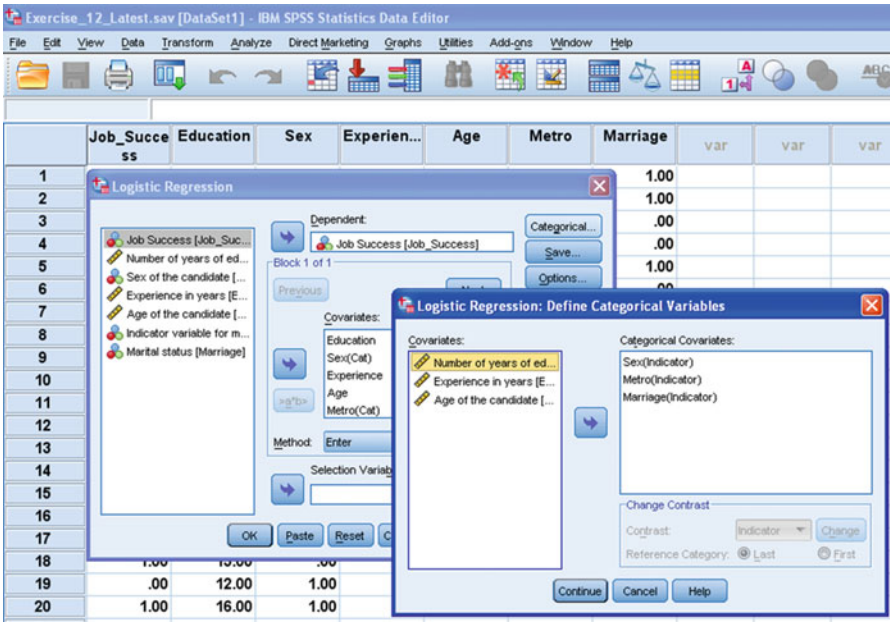


Fig. 13.6 Screen showing selection of variables for logistic regression

The Forward:LR method is considered to be the most efficient method among different forward and backward methods. Here LR refers to the likelihood ratio method. In this method, the variables are selected in the model one by one based on their utility. To select this method in the SPSS, do the following steps:

- Select the option "Forward:LR" by using the dropdown menu of the command **Method** in the screen shown in Fig. 13.6.
- Click **OK.**

(vi) *Getting the output:* Clicking the option **OK** shall generate lots of output in the output window. These outputs may be selected from the output window by using right click of the mouse and may be copied in the word file. The relevant outputs so selected for discussion are shown in Tables 13.3–13.12. One must understand the meaning of these outputs so that while writing thesis or project report, they may be incorporated with proper explanation.

Interpretation of Various Outputs Generated in Logistic Regression

Descriptive Findings

Table 13.3 shows the number of cases (*N*) in each category (e.g., included in the analysis, missing, and total) and their percentage. In logistic regression, a listwise deletion of missing data is done by default in SPSS. Since there is no missing data,

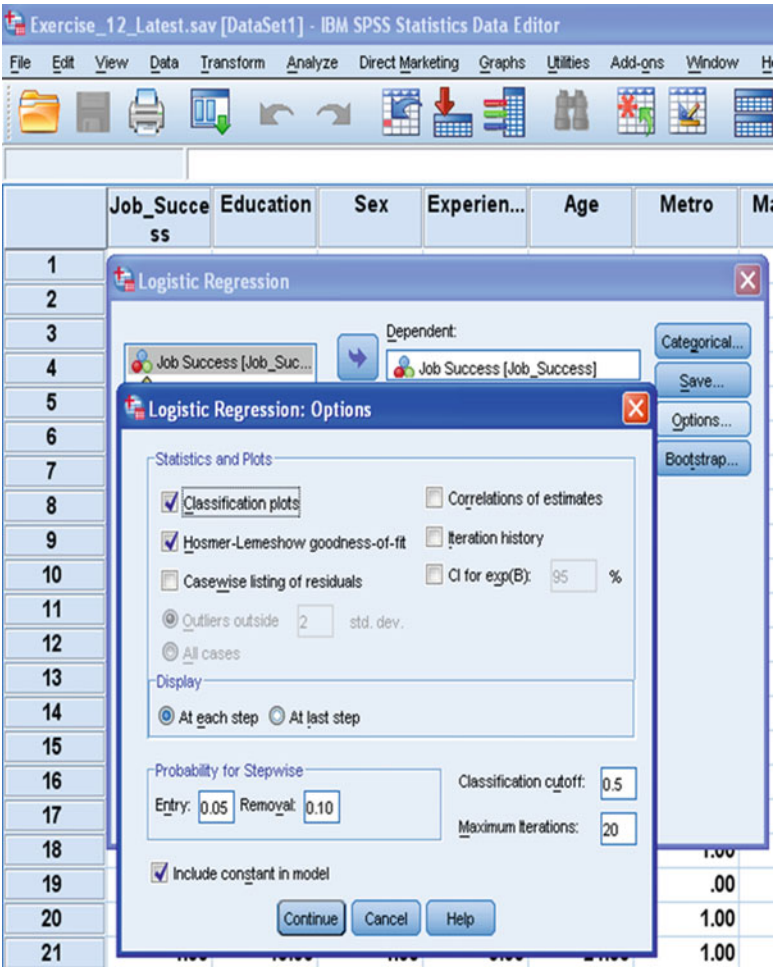


Fig. 13.7 Screen showing option for generating classification plots and Hosmer-Lemeshow goodness-of-fit

the number of missing cases is shown as 0. Table 13.4 shows the coding of the dependent variable used in the data file, that is, 1 for success and 0 for failure in getting the job.

Table 13.5 shows the coding of all the categorical independent variables along with their frequencies in the study. While coding the categorical variables, highest number should be allotted to the reference category because by default SPSS considers the category with the highest coding as the reference category and gives the code as 0. For instance, if you define the coding of the variable sex as 0 for "female" and 1 for "male," then the SPSS will consider male as the reference category and convert its code to 0 and the other category female as 1.

Table 13.3 Case processing summary

Unweighted cases[a]		*N*	Percent
Selected cases	Included in analysis	23	100.0
	Missing cases	0	.0
	Total	23	100.0
Unselected cases		0	.0
Total		23	100.0

[a]If weight is in effect, see classification table for the total number of cases

Table 13.4 Dependent variable encoding

Original value	Internal value
Failure	0
Success in getting job	1

Table 13.5 Categorical variable coding

			Parameter coding
		Frequency	(1)
Marital status	Unmarried	10	1.000
	Married	13	.000
Metro	Nonmetro	11	1.000
	Metro	12	.000
Sex	Female	7	1.000
	Male	16	.000

If you look into the coding of the independent categorical variables in the Table 13.2, that is, sex (0:female, 1:male), metro (0:nonmetro resident, 1:metro resident), and marital status (0:unmarried and 1:married), these coding have been reversed by the SPSS as shown in the Table 13.5. It is because SPSS by default considers the highest coding as the reference category and converts it into 0. However, you can change the reference category as the lowest coding in SPSS screen as shown in the Fig. 13.6.

Analytical Findings

The findings in this section are the most interesting part of the output. These findings include the test of the overall model, significance of regression coefficients, and the values of the odds ratios.

In this study, since the **Forward:LR** method has been chosen for logistic regression, you will get more than one model with different number of variables in it. The results of the logistic regression shall be discussed in two blocks. In the first block, the logistic regression model shall be developed by using the constant without using any of the independent variables. This model may be used to compare the utility of the model developed in block two by using the identified independent variables.

Table 13.6 Classification table[a, b]

			Predicted		
			Job		Percentage correct
Observed			Failure	Success	
Step 0	Job	Failure	0	10	0
		Success	0	14	100.0
		Overall Percentage			58.3

[a]Constant is included in the model
[b]The cut value is .500

Table 13.7 Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(*B*)
Step 0	Constant	.336	.414	.660	1	.416	1.400

Block 0: Beginning Block

In Block 0, the results are shown for the model with only the constant included before any coefficients (i.e., those relating to education, sex, experience, age, metro, and marital) are entered into the equation. Logistic regression compares the model obtained in Block 0 with a model including the predictors to determine whether the latter model is more efficient. The Table 13.6 shows that if nothing is known about the independent variables and one simple guesses that a person would be selected for the job, we would be correct 58.3% of the time. Table 13.7 shows that the Wald statistics is not significant as its significance value is 0.416 which is more than 0.05. Hence, the model with constant is not worth and is equivalent to just guessing about the target variable in the absence of any knowledge about the independent variables.

Table 13.8 shows whether each independent variable improves the model or not. You can see that the variables sex, metro, and marital may improve the model as they are significant with sex and marital slightly better than metro. Inclusion of these variables would add to the predictive power of the model. If these variables had not been significant and able to contribute to the prediction, then the analysis would obviously be terminated at this stage.

Block 1 Method = Forward:LR

In this block, results of the different models with different independent variables shall be discussed.

Table 13.9 shows the value of −2 log likelihood (−2LL), which is a deviance statistic between the observed and predicated values of the dependent variable. If this deviance statistic is insignificant, it indicates that the model is good and there is no difference between observed and predicted values of dependent variable. This number in absolute term is not very informative. However, it can be used to compare different models having different number of predictive variables. For

Table 13.8 Variables not in the equation

			Score	df	Sig.
Step 0	Variables	Education	1.073	1	.300
		Sex(1)	7.726	1	.005
		Experience	.728	1	.393
		Age	.174	1	.677
		Metro(1)	4.608	1	.032
		Marital(1)	8.061	1	.005
		Overall statistics	14.520	6	.024

Table 13.9 Model summary

Step	−2 log likelihood	Cox and Snell *R*-square	Nagelkerke *R*-square
1	24.053[a]	.300	.403
2	18.549[b]	.443	.597

[a]Estimation terminated at iteration number 4 because parameter estimates changed by less than .001

[b]Estimation terminated at iteration number 5 because parameter estimates changed by less than .001

instance, in Table 13.9, the value of −2LL has reduced from 24.053 to 18.549. This indicates that there is an improvement in model 2 by including an additional variable, sex. In fact, the value of −2LL should keep on decreasing if you go on adding the significant predictive variables in the model.

Unlike OLS regression equation, there is no concept of R^2 in logistic regression. It is because of the fact that the dependent variable is dichotomous and R^2 cannot be used to show the efficiency of prediction. However, several authors have suggested pseudo *R*-squares which are not equivalent to the *R*-square that is calculated in OLS regression. Thus, this statistic should be interpreted with great caution. Two such pseudo *R*-squares suggested by Cox and Snell and Nagelkerke are shown in Table 13.9. As per Cox and Snell's R^2, 44.3% of the variation in the dependent variable is explained by the logistic model. On the other hand, Nagelkerke's R^2 explains 59.7% variability of the dependent variable by the independent variables in the model. Nagelkerke's R^2 is more reliable measure of relationship in comparison to Cox and Snell's R^2. Nagelkerke's R^2 will normally be higher than Cox and Snell's R^2.

In order to find whether the deviance statistic −2 log likelihood is insignificant or not, Hosmer and Lemeshow suggested the chi-square statistic which is shown in Table 13.10. In order that the model is efficient, this chi-square statistic should be insignificant. Since the *p* value associated with chi-square in Table 13.10 is .569 for the second model, which is greater than .05, it is insignificant and it can be interpreted that the model is efficient.

Table 13.11 is a classification table which shows the observed and predicted values of the dependent variable in both the models. In the second model, it can be seen that out of 10 candidates who did not get the success in getting the job, four were wrongly predicted to get the job. Similarly out of 14 candidates who succeeded to get the job, none was wrongly predicted to be failure. Thus, the model correctly classified 83.3% cases. This can be obtained by (20/24) × 100.

Table 13.10 Hosmer and Lemeshow test

Step	Chi-square	df	Sig.
1	.000	0	.
2	1.129	2	.569

Table 13.11 Classification table[a]

			Predicted		
			Job		Percentage correct
Observed			Failure	Success	
Step 1	Job	Failure	8	2	80.0
		Success	3	11	78.6
		Overall percentage			79.2
Step 2	Job	Failure	6	4	60.0
		Success	0	14	100.0
		Overall percentage			83.3

[a]The cut value is .500

Table 13.12 Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(*B*)
Step 1[a]	Marital(1)	−2.686	1.024	6.874	1	.009	.068
	Constant	1.705	.769	4.918	1	.027	5.500
Step 2[b]	Sex(1)	−2.666	1.278	4.352	1	.037	.070
	Marital(1)	−2.711	1.253	4.682	1	.030	.066
	Constant	2.779	1.146	5.886	1	.015	16.106

[a]Variable(s) entered on step 1: marital
[b]Variable(s) entered on step 2: sex

Table 13.12 is the most important table which shows the value of regression coefficients *B*, Wald statistics, its significance, and odds ratio exp(*B*) for each variable in both the models. The *B* coefficients are used to develop the logistic regression equation for predicting the dependent variable from the independent variables. These coefficients are in log-odds units. Thus, the logistic regression equation in the second model is given by $\log \frac{p}{1-p} = 2.779 - 2.666 \times \text{Sex}(1) - 2.711 \times \text{Marital}(1)$where *p* is the probability of getting the job. The dependent variable in the logistic regression is known as logit(*p*) which is equal to $\log(p/(1 - p))$.

The estimates obtained in the above logistic regression equation explain the relationship between the independent variables and the dependent variable, where the dependent variable is on the logit scale. These estimates tell the amount of increase (or decrease, if the sign of the coefficient is negative) in the estimated log odds of "job success" = 1 that would be predicted by a 1 unit increase (or decrease) in the predictor, holding all other predictors constant.

Because regression coefficients *B* are in log-odds units, they are often difficult to interpret; hence, they are converted into odds ratios which are equal to exp(*B*). These odds ratios are shown in the last column of Table 13.12.

Significance of the Wald statistics indicates that the variable significantly predicts the success in getting the bank job, but it should be used only in a situation

where the sample size is quite large, preferably more than 500. In case of small sample, the level of significance gets inflated and it does not give the correct picture. Since in this problem the value of chi-square in Hosmer and Lemeshow test as shown in Table 13.10 is insignificant, the model can be considered to be valid for predicting the success in getting the bank's job on the basis of the second model with two independent variables, that is, marital and sex.

Explanation of Odds Ratio

In Table 13.12, the exp(*B*) represents the odds ratio for all the predictors. If the value of the odds ratio is large, its predictive value is also large. Since the second model is the final model in this study, the discussion shall be done for the variables in this model only. Here both the independent variables, that is, sex and marital, are significant. Since the sex(1) variable has a larger odds ratio .070, this is slightly a better predictor in comparison to marital(1) variable in getting the bank's job.

The value of exp(*B*) for the variable sex(1) is 0.070. It indicates that if the candidate appearing in the bank exams is female, then there would be decrease in the odds of 93% (.07–1.00 = −.93). In other words, if a female candidate is appearing in the bank examination, her chances of success would be 93% less than the men candidate if other variables are kept constant. Similarly the exp(*B*) value of the variable marital(1) is .066. This indicates that there would be decrease in the odds of 93.4% (.066–1.000 = −.934). It can be interpreted that if the candidate appearing in the bank examination is unmarried, his/her chances of success would be 93.4% less than the married candidate provided other variables are kept constant.

Conclusion

To conclude, if the candidate is male and married, the chances of odds increases for getting selected for a bank job in comparison to female and unmarried candidate.

Summary of the SPSS Commands for Logistic Regression

(i) Start SPSS and prepare the data file by defining the variables and their properties in **Variable View** and typing the data column-wise in Data View.
(ii) In the Data View, follow the below-mentioned command sequence for factor analysis:

Analyze ⟶ Regression ⟶ Binary Logistic

(iii) Select the dependent variable from the left panel to the "Dependent" section in the right panel and all independent variables including categorical variables from left panel to the "Covariates" section in the right panel.
(iv) By clicking the **Categorical command**, select the categorical variables from the "Covariates" section to the "Categorical Covariates" in the right panel and click ***Continue.***
(v) Click the tag **Options** and check "Classification Plots" and "Hosmer-Lemeshow goodness-of-fit" and click ***Continue***.
(vi) Ensure that the option **Forward:LR** is chosen by default and then click **OK** for output.

Exercise

Short Answer Questions

Note: Write answer to each questions in not more than 200 words.

Q.1. What is logit and how is it used to interpret the probability of success?
Q.2. What do you mean by odds ratio? Explain the monotonic transformation in relation with odds ratio and log odds.
Q.3. Explain the logistic function and its characteristics.
Q.4. Why is the logit function used in logistic regression analysis?
Q.5. Explain the meaning of maximum likelihood and the significance of -2 log likelihood.
Q.6. What is the difference between logic regression and OLS regression?
Q.7. How are the dummy variables created in a situation where an independent categorical variable has more than two options?
Q.8. Write any four assumptions used in logistic regression.
Q.9. What are the advantages of using logistic regression analysis?
Q.10. Explain any one research situation in detail where logistic regression can be applied.
Q.11. Write in brief the various steps involved in logistic regression.
Q.12. What is Hosmer and Lemeshow test? How is it used and what does it indicate?

Multiple-Choice Questions

Note: For each of the question, there are four alternative answers for each question. Tick mark the one that you consider the closest to the correct answer.

1. Logistic regression is used when the dependent variable is

   (a) Continuous
   (b) Ordinal
   (c) Binary
   (d) Categorical

2. If $\exp(3) = 20.12$, then $\log(20.12)$ is

   (a) 20.12
   (b) 23.12

(c) 17.12
(d) 3

3. If the probability of success is 0.6, then the odds of success is

(a) 0.4
(b) 1.5
(c) 2.4
(d) 0.75

4. In a logistic regression, if the odds ratio for an independent variable is 2.5, then which of the following is true?

(a) The probability of the dependent variable happening is 0.25.
(b) The odds against the dependent variable happening is 2.5.
(c) The odds for the dependent variable happening is 2.5.
(d) The odds for the dependent variable happening is 2.5 against one unit increase in the independent variable.

5. If p is the probability of success, then the logit of p is

(a) $\ln \frac{1-p}{p}$
(b) $\ln \frac{1+p}{p}$
(c) $\log \frac{p}{1-p}$
(d) $\log \frac{p}{1+p}$

6. The logistic function $f(z)$ is equal to

(a) $\frac{e^z}{1+e^z}$
(b) $\frac{1+e^z}{e^z}$
(c) $\frac{e^z}{1-e^z}$
(d) $\frac{1-e^z}{e^z}$

7. In logistic regression, odds ratio is equivalent to

(a) Log(B)
(b) Exp(B)
(c) B coefficient
(d) $\frac{p}{1-p}$

8. Choose the correct statement.

(a) The independent variable is required to be linearly related with the dependent variable.
(b) The independent variable is required to be linearly related with logit transformation of the outcome variable.
(c) The dependent variable is always continuous.
(d) Probability of success in the outcome variable is equivalent to the log odds.

9. Choose the correct command for starting logistic regression in SPSS.

   (a) Analyze → Regression → Binary Logistic
   (b) Analyze → Regression → Logistic Regression
   (c) Analyze → Binary Logistic → Regression
   (d) Analyze → Logistic → Binary Regression

10. In using the Hosmer-Lemeshow goodness-of-fit, model is considered to be good if

    (a) Chi-square is significant at any predefined level.
    (b) Chi-square is not significant at any predefined level.
    (c) Chi-square is equal to 100.
    (d) All the regression coefficients are significant.

Assignments

1. Following are the scores of 90 candidates in different subjects obtained in a MBA entrance examination. Apply the logistic regression to develop a model for predicting success in the examination on the basis of independent variables. Discuss the comparative importance of independent variables in predicting success in the examination. For the variable MBA, coding 1 represents success and 0 indicates failure in the examination. Similarly gender 1 indicates male and 2 indicates female.

MBA	English	Reasoning	Math	Gender	MBA	English	Reasoning	Math	Gender
1	68	50	65	0	0	46	52	55	1
0	39	44	52	1	0	39	41	33	0
0	44	44	46	1	0	52	49	49	0
1	50	54	61	1	0	28	46	43	0
1	71	65	72	0	0	42	54	50	1
1	63	65	71	1	0	47	42	52	0
0	34	44	40	0	0	47	57	48	1
1	63	49	69	0	0	52	59	58	0
0	68	43	64	0	0	47	52	43	1
0	47	45	56	1	1	55	62	41	0
0	47	46	49	1	0	44	52	43	0
0	63	52	54	0	0	47	41	46	0
0	52	51	53	0	0	45	55	44	1
0	55	54	66	0	0	47	37	43	0
1	60	68	67	1	0	65	54	61	0
0	35	35	40	0	0	43	57	40	1
0	47	54	46	1	0	47	54	49	0
1	71	63	69	0	1	57	62	56	0
0	57	52	40	1	0	68	59	61	1
0	44	50	41	0	0	52	55	50	0
0	65	46	57	0	0	42	57	51	0
1	68	59	58	1	0	42	39	42	1

(continued)

MBA	English	Reasoning	Math	Gender	MBA	English	Reasoning	Math	Gender
1	73	61	57	1	1	66	67	67	1
0	36	44	37	0	1	47	62	53	0
0	43	54	55	0	0	57	50	50	0
1	73	62	62	1	1	47	61	51	1
0	52	57	64	1	1	57	62	72	1
0	41	47	40	0	0	52	59	48	1
0	50	54	50	0	0	44	44	40	1
0	50	52	46	1	0	50	59	53	1
0	50	52	53	0	0	39	54	39	0
0	47	46	52	0	1	57	62	63	1
1	62	62	45	1	0	57	50	51	1
0	55	57	56	1	0	42	57	45	0
0	50	41	45	1	0	47	46	39	0
0	39	53	54	1	0	42	36	42	1
0	50	49	56	0	0	60	59	62	0
0	34	35	41	0	0	44	49	44	0
0	57	59	54	1	0	63	60	65	1
1	65	60	72	0	1	65	67	63	1
1	68	62	56	0	0	39	54	54	0
0	42	54	47	0	0	50	52	45	1
0	53	59	49	1	1	52	65	60	0
1	59	63	60	1	1	60	62	49	1
0	47	59	54	1	0	44	49	48	0

2. In an assembly election, victory of a candidate depends upon many factors. In order to develop a model for predicting the success of a candidate (1 if elected and 0 if not elected) on the basis of independent variables, the data on 30 contestants were obtained on the variables like candidate's age, sex (1 for male and 0 for female), experience in politics, status in politics (1 for full time and 0 for part time), education (in number of years), and elected history (1 if elected earlier

Profile data of the contestants in the assembly election

Election result	Age (in years)	Sex	Experience (in years)	Status in politics	Education (no. of years)	Election history
1.00	48.00	1.00	10.00	1.00	15.00	1.00
1.00	42.00	1.00	16.00	1.00	18.00	1.00
1.00	46.00	1.00	12.00	1.00	15.00	.00
.00	42.00	.00	16.00	.00	16.00	1.00
.00	45.00	.00	20.00	1.00	18.00	1.00
1.00	47.00	1.00	18.00	.00	15.00	.00
.00	34.00	.00	28.00	.00	15.00	.00
.00	47.00	1.00	20.00	.00	12.00	1.00
.00	36.00	1.00	30.00	1.00	10.00	1.00
1.00	63.00	.00	35.00	.00	16.00	.00
.00	45.00	1.00	25.00	1.00	12.00	.00
1.00	54.00	.00	20.00	1.00	16.00	1.00

(continued)

Election result	Age (in years)	Sex	Experience (in years)	Status in politics	Education (no. of years)	Election history
1.00	58.00	1.00	34.00	.00	18.00	.00
.00	54.00	.00	38.00	.00	12.00	.00
.00	56.00	1.00	35.00	.00	10.00	1.00
1.00	55.00	.00	30.00	1.00	15.00	.00
1.00	54.00	.00	31.00	1.00	16.00	.00
1.00	58.00	1.00	34.00	.00	15.00	1.00
.00	37.00	1.00	35.00	.00	10.00	.00
1.00	45.00	1.00	22.00	.00	15.00	1.00
.00	34.00	1.00	5.00	1.00	12.00	1.00
.00	47.00	.00	9.00	.00	12.00	1.00
.00	42.00	1.00	8.00	.00	12.00	1.00
1.00	45.00	1.00	6.00	.00	15.00	.00
1.00	28.00	1.00	2.00	1.00	16.00	1.00
1.00	43.00	.00	12.00	1.00	16.00	1.00
.00	35.00	1.00	11.00	1.00	15.00	1.00
1.00	43.00	.00	18.00	1.00	15.00	.00
1.00	45.00	.00	17.00	.00	16.00	1.00
1.00	41.00	1.00	13.00	.00	15.00	1.00
.00	42.00	1.00	15.00	1.00	15.00	.00

and 0 if not elected earlier). Apply the logistic regression and develop the model for predicting success in assembly election.

Answers to Multiple-Choice Questions

Q.1 c Q.2 d Q.3 b Q.4 d
Q.5 c Q.6 a Q.7 b Q.8 b
Q.9 a Q.10 b