CHAPTER 11

LOGISTIC REGRESSION

In Chapter 11, we present discussion of logistic regression—an alternative to discriminant analysis, as well as multiple regression in certain situations. Logistic regression has the same basic purpose as discriminant analysis—the classification of individuals into groups. It is, in some ways, more flexible and versatile than discriminant analysis, although mathematically, it can be quite a bit more cumbersome. In this chapter, we present discussion of logistic regression, which seeks to identify a combination of IVs—which are limited in few, if any, ways—that best predicts membership in a particular group, as measured by a categorical DV.

SECTION 11.1 PRACTICAL VIEW

Purpose

Logistic regression is basically an extension of multiple regression in situations where the DV is not a continuous or quantitative variable (George & Mallery, 2000). In other words, the DV is categorical (or discrete) and may have as few as two values. For example, in a logistic regression application, these categories might include values such as membership or nonmembership in a group, completion or noncompletion of an academic program, passing or failure to pass a course, survival or failure of a business, etc.

Due to the nature of the categorical DV in logistic regression, this procedure is also sometimes used as an alternative to discriminant analysis. Since the goal is to predict values on a DV that is categorical, we are essentially attempting to predict membership into one of two or more "groups." The reader should likely see the similarity between this procedure and that discussed in the previous chapter for discriminant analysis (i.e., the classification, or prediction, of subjects into groups). Although logistic regression may be used to predict values on a DV of two or more categories, our discussion will focus on binary logistic regression, in which the DV is dichotomous.

The basic concepts fundamental to multiple regression analysis—namely that several variables are regressed onto another variable using one of several selection processes—are the same for logistic regression analysis (George & Mallery, 2000), although the meaning of the resultant regression equation is considerably different. As you read in Chapter 7, a standard regression equation is composed of the sum of the products of weights and actual values on several predictor variables (IVs) in order to predict the values on the criterion variable (DV). In contrast, the value that is being predicted in logistic regression is actually a *probability*, which ranges from 0 to 1. More precisely, logistic regression specifies the probabilities of the particular outcomes (e.g., "pass" and "fail") for each subject or case involved. In other words, logistic regression analysis produces a regression equation that accurately predicts the probability of whether an individual will fall into one category (e.g., "pass") or the other (e.g., "fail") (Tate, 1992).

Although logistic regression is fairly similar to both multiple regression and discriminant analysis, it does provide several distinct advantages over both techniques (Tabachnick & Fidell, 1996). Unlike both discriminant analysis and multiple regression, logistic regression requires that no assumptions about the distributions of the predictor variables (IVs) need to be made by the researcher. In other words, the predictors do not have to be normally distributed, linearly related, or have equal variances within each group. It should be obvious to the reader that this fact alone makes logistic regression much more flexible than the other two techniques. Additionally, logistic regression cannot produce negative predictive probabilities, as can happen when applying multiple regression to situations involving dichotomous outcomes (Tate, 1992). In logistic regression, all probability values will be positive and will range from 0 to 1 (Tabachnick & Fidell, 1996). Another advantage is that logistic regression has the capacity to analyze predictor variables of all types—continuous, discrete, and dichotomous. Finally, logistic regression can be especially useful when the distribution of data on the criterion variable (DV) is expected or known to be *nonlinear* with one or more of the predictor variables (IVs). Logistic regression is able to produce nonlinear models, which again adds to its overall flexibility.

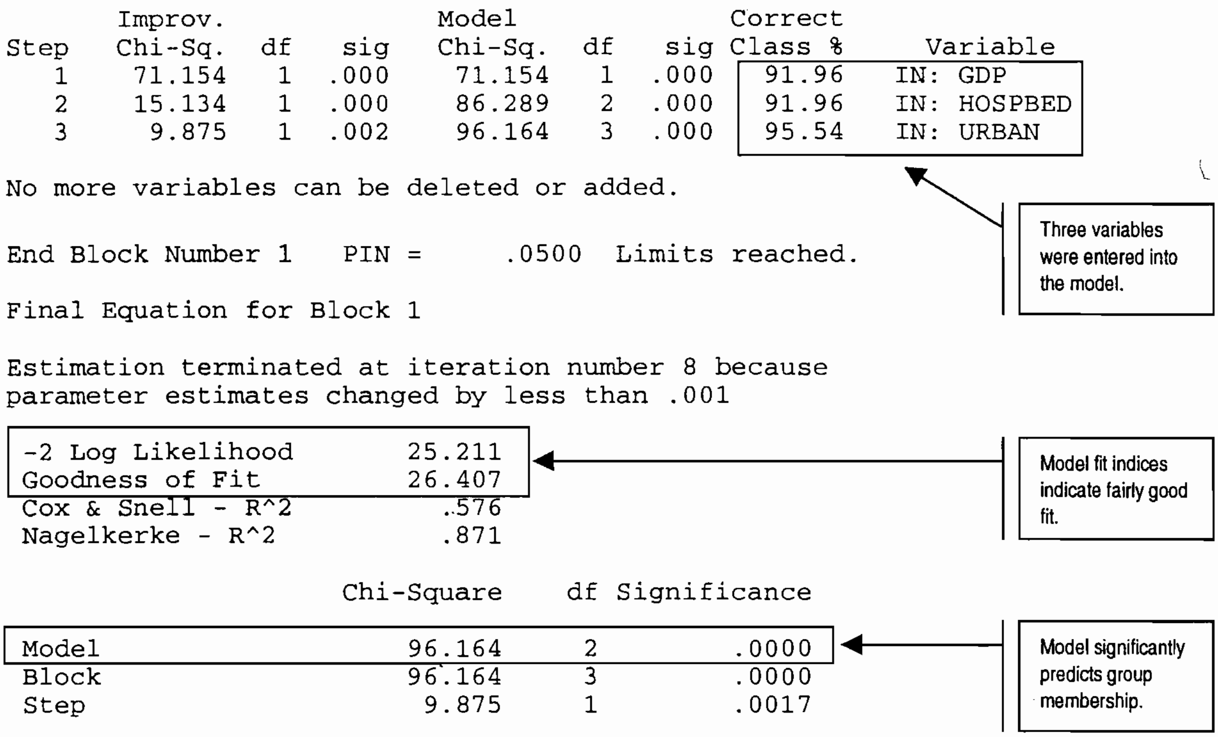
Let us develop an initial working example to which we will refer in this chapter. Suppose we wanted to determine a nation's status in terms of its level of development based on several measures. Specifically, those measures consist of:

- population (1992) in millions (*pop92*);
- percent of the population living in urban areas (*urban*);
- gross domestic product per capita (*gdp*);
- death rate per 1,000 people (*deathrat*);
- number of radios per 100 people (*radio*);
- number of hospital beds per 10,000 people (*hospbed*); and
- number of doctors per 10,000 (*docs*).

Combinations of these seven variables that would accurately predict the probability of a country's status as a developing nation (*develop*)—either (1) developed nation or (2) developing nation—will be determined as a result of a logistic regression analysis. Our sample analysis was conducted using a forward method of entry for the seven predictors.

The results obtained from a logistic regression analysis are somewhat different from those that we have seen accompany previous analysis techniques. There are basically three main output components to be interpreted. First, the resulting model is evaluated using goodness-of-fit tests. The table showing the results of chi-square goodness-of-fit tests for our working example is presented in Figure 11.1. One should notice that the model resulted in the inclusion of three variables from the original seven predictors—*gdp* (entered in Step 1), *hospbed* (entered in Step 2), and *urban* (entered at Step 3). At each step, this test essentially compares the actual values for cases on the DV with the predicted values on the DV. All steps resulted in significance values < .001, indicating that these three variables are significant and important predictors of the DV, *develop*. Also included in this table are the percentages of correct classification—based on the model—at each step (i.e., based on the addition of each variable). In Step 1, the model with only *gdp* correctly classified 91.96% of the cases; in Step 2, the model with *gdp* and *hospbed* correctly classified 91.96%; in Step 3, the model of *gdp, hospbed*, and *urban* correctly classified 95.54% of the cases.

Figure 11.1 Goodness-of-fit Indices for Example Number 1.



```
        Improv.                  Model              Correct
Step    Chi-Sq.   df    sig    Chi-Sq.   df    sig  Class %      Variable
  1     71.154    1   .000     71.154    1   .000    91.96     IN: GDP
  2     15.134    1   .000     86.289    2   .000    91.96     IN: HOSPBED
  3      9.875    1   .002     96.164    3   .000    95.54     IN: URBAN

No more variables can be deleted or added.

End Block Number 1    PIN =      .0500  Limits reached.

Final Equation for Block 1

Estimation terminated at iteration number 8 because
parameter estimates changed by less than .001

 -2 Log Likelihood          25.211
 Goodness of Fit            26.407
 Cox & Snell - R^2            .576
 Nagelkerke - R^2             .871

                         Chi-Square    df  Significance

 Model                      96.164      2          .0000
 Block                      96.164      3          .0000
 Step                        9.875      1          .0017
```

Three variables were entered into the model.

Model fit indices indicate fairly good fit.

Model significantly predicts group membership.

The second table shown in Figure 11.1 includes several indices of *overall* model fit. Smaller values on the first measure, labeled ***-2 Log Likelihood***, indicate that the model fits the data better; a perfect model has a value for this measure equal to 0 (George & Mallery, 2000). The next measure, ***Goodness-of-Fit***, compares the actual values for cases on the DV with the predicted values on the DV; this measure is similar to the chi-square value in the first table. The third and fourth measures, ***Cox & Snell – R^2*** and ***Nagelkerke – R^2***, are essentially estimates of R^2 indicating the proportion of variability in the DV that may be accounted for by all predictor variables included in the equation.

The second component is a classification table for the DV. The classification table for *develop* is presented in Figure 11.2. The classification table compares the predicted values for the DV, based on the logistic regression model, with the actual observed values from the data. The predicted values are obtained by computing the probability for a particular case (this computation will be discussed later in Section 11.3) and classifies it into one of the two possible categories based on that probability. If the calculated probability is less than .50, the case is classified into the first value on the DV—in our example, the first category is *developed* nation (coded "0").

The third and final component to be interpreted is the table of coefficients for variables included in the model. The coefficients for our working example are shown in Figure 11.3. These coefficients are interpreted in similar fashion to coefficients resulting from a multiple regression. The values labeled *B* are the regression coefficients or weights for each variable used in the equation. The significance of each predictor is tested not with a *t*-test, as in multiple regression, but with a measure known as the ***Wald statistic*** and the associated significance value. The value *R* is the partial correlation coefficient between each predictor variable and the DV, holding constant all other predictors in the equation. Finally, ***Exp(B)*** provides an alternative method of interpreting the regression coefficients. The meaning of this coefficient will be explained further in Section 11.3.

Figure 11.2 Classification Table for Example Number 1.

```
Classification Table for DEVELOP
The Cut Value is .50
                                          Predicted
                              Developed countr Developing count  Percent Correct
                                             0                1
                                     +----------------+----------------+
Observed
   Developed countr    0       |      23        |        3       |    88.46%
                                     +----------------+----------------+
   Developing count    1       |       2        |       84       |    97.67%
                                     +----------------+----------------+
                                                         Overall   95.54%
```

Model is extremely accurate in classifying subjects.

Figure 11.3 Regression Coefficients for Example Number 1.

```
---------------------- Variables in the Equation -----------------------

Variable            B       S.E.      Wald     df      Sig        R     Exp(B)

URBAN           .1354      .0671    4.0782      1    .0434    .1309     1.1450
GDP            -.0010      .0004    6.6044      1    .0102   -.1948      .9990
HOSPBED        -.0832      .0271    9.4113      1    .0022   -.2471      .9202
Constant       3.0472     1.2516    5.9272      1    .0149
```

Odds ratios are fairly small.

Sample Research Questions

The goal of logistic regression analysis is to correctly predict the category of outcome for individual cases. Further, attempts are made to reduce the number of predictors (in order to achieve parsimony) while maintaining a strong level of prediction. Based on our working example, let us now proceed to the specification of a series of possible research questions for our analysis:

(1) Can status as a developing nation (i.e., develop*ed* or develop*ing*) be correctly predicted from knowledge of population; percent of population living in urban areas; gross domestic product; death rate; number of radios, hospital beds, and doctors?

(2) If developing nation status can be predicted correctly, which variables are central in the prediction of that status? Does the inclusion of a particular variable increase or decrease the probability of the specific outcome?

(3) How good is the model at classifying cases for which the outcome is unknown? In other words, how many developing countries are classified correctly? How many developed countries are classified correctly?

Section 11.2 Assumptions and Limitations

As mentioned earlier, logistic regression does not require the adherence to any assumptions about the distributions of predictor variables (Tabachnick & Fidell, 1996). However, if distributional assumptions are met, discriminant analysis may be a stronger analysis technique; thus, the researcher may want to opt for this procedure.

There are, however, several important issues related to the use of logistic regression. First, there is the issue of the ratio of cases to variables included in the analysis. Several problems may occur if too few cases relative to the number of predictor variables exist in the data. Logistic regression may produce extremely large parameter estimates and standard errors, especially in situations where combinations of discrete variables result in too many cells with no cases. If this situation occurs, the researcher is advised to collapse categories of the discrete variables, delete any offending categories (if patterns are evident), or simply delete any discrete variable if it is not important to the analysis (Tabachnick & Fidell, 1996). Another option open to the researcher is to increase the number of cases in the hope that this will "fill in" some of the empty cells.

Second, logistic regression relies on a goodness-of-fit test as a means of assessing the fit of the model to the data. You may recall from an earlier course in statistics that a goodness-of-fit test includes values for the expected frequencies for each cell in the data matrix formed by combinations of discrete variables. If any of the cells have expected frequencies that are too small (typically, $f_e < 5$), the analysis may have little power (Tabachnick & Fidell, 1996). All pairs of discrete variables should be evaluated to ensure that all cells have expected frequencies greater than 1 and that no more than 20% have frequencies less than 5. If either of these conditions fail, the researcher should consider accepting a lower level of power for the analysis, collapsing categories for variables with more than two levels, or deleting discrete variables so as to reduce the total number of cells (Tabachnick & Fidell, 1996).

Third, as with all varieties of multiple regression, logistic regression is sensitive to high correlations among predictor variables. This condition results in multicollinearity among predictor variables, as discussed in Chapter 7. If multicollinearity is present among variables in the analysis, one is advised to delete one or more of the redundant variables from the model in order to eliminate the multicollinear relationships (Tabachnick & Fidell, 1996).

Finally, extreme values on predictor variables should be examined carefully. As with multiple regression, resultant logistic regression models are very sensitive to outliers. A case that is actually in one outcome category may show a high probability for being in another category. Multiple cases such as this will result in a model with extremely poor fit. Standardized residuals should be examined in order to detect outliers; any identified outliers—those cases with values > |3|—should be addressed using standard methods (i.e., deletion from the sample).

Section 11.3 Process and Logic

The Logic Behind Logistic Regression

Mathematically speaking, logistic regression is based on probabilities, odds, and the logarithm of the odds (George & Mallery, 2000). Probabilities are simply the number of outcomes of a specific type expressed as a proportion of the total number of possible outcomes. For example, if we roll a single die, the probability of rolling a three would be 1 out of 6—there is only one "3" on a die and there are six possible outcomes. This ratio could also be expressed as a proportion (.167) or as a percentage (16.7%). In a logistic regression application, ***odds*** are defined as the ratio of the probability that an event will occur divided by the probability that the event will not occur. In other words,

$$Odds = \frac{p(X)}{1 - p(X)} \qquad \text{(Equation 11.1)}$$

where *p(X)* is the probability of event *X* occurring and *1-p(X)* is the probability of event *X* not occurring. Therefore, the odds of rolling a "3" on a die are

$$Odds_{"3"} = \frac{p("3")}{1 - p("3")} = \frac{.167}{.833} = .200$$

It is important for the reader to keep in mind that probabilities will always have values that range from 0 to 1, but odds may be greater than 1. Applying the concept of odds to our working logistic regression example of classification as a developing nation would give us the following equation:

$$Odds_{developing} = \frac{p(developing)}{1 - p(developing)}$$

The effect of an IV on a dichotomous outcome is usually represented by an odds ratio. The ***odds ratio***—symbolized by ψ or Exp(B)—is defined as a ratio of the odds of being classified in one category (i.e., Y=0 or Y=1) of the DV for two different values of the IV (Tate, 1992). For example, we would be interested in the odds ratio for being classified as a "developing nation" (Y=0) for a given increase in the value of the score on the combination of the three significant predictors of *develop*, namely *urban, gdp* and *hospbed.*

The ultimate model obtained by a logistic regression analysis is a nonlinear function (Tate, 1992). A key concept in logistic regression with which the reader must be familiar is known as a logit. A ***logit*** is the natural logarithm of the odds—an operation that most pocket calculators will perform. Again extending our simplified example, the logit for our odds of rolling a "3" would be

$$\ln(.200) = -1.609$$

Specifically, in logistic regression, $\hat{Y}$ is the probability of having one outcome or another based on a nonlinear model resulting from the best linear combination of predictors. We can combine the ideas of probabilities, odds, and logits into one equation:

$$\hat{Y}_i = \frac{e^u}{1 + e^u} \qquad \text{(Equation 11.2)}$$

where $\hat{Y}_i$ is the estimated probability that the i^{th} case is in one of the categories of the DV, and e is a constant equal to 2.718, raised to the power u, where u is the usual regression equation:

$$u = B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k \qquad \text{(Equation 11.3)}$$

The linear regression equation (u) is then the natural log of the probability of being in one group divided by the probability of being in the other group (Tabachnick & Fidell, 1996). The linear regression equation creates the logit or log of the odds:

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = B_0 + B_1X_1 + B_2X_2 + \ldots + \mathrm{B}_k\mathrm{X}_k \qquad \text{(Equation 11.4)}$$

We tend to agree with George & Mallery (2000), who state in their text, "[These] equation[s are] probably not very intuitive to most people...it takes a lot of experience before interpreting logistic regression equations becomes intuitive." For most researchers, focus should more appropriately be directed at assessing the fit of the model, as well as its overall predictive accuracy.

Interpretation of Results

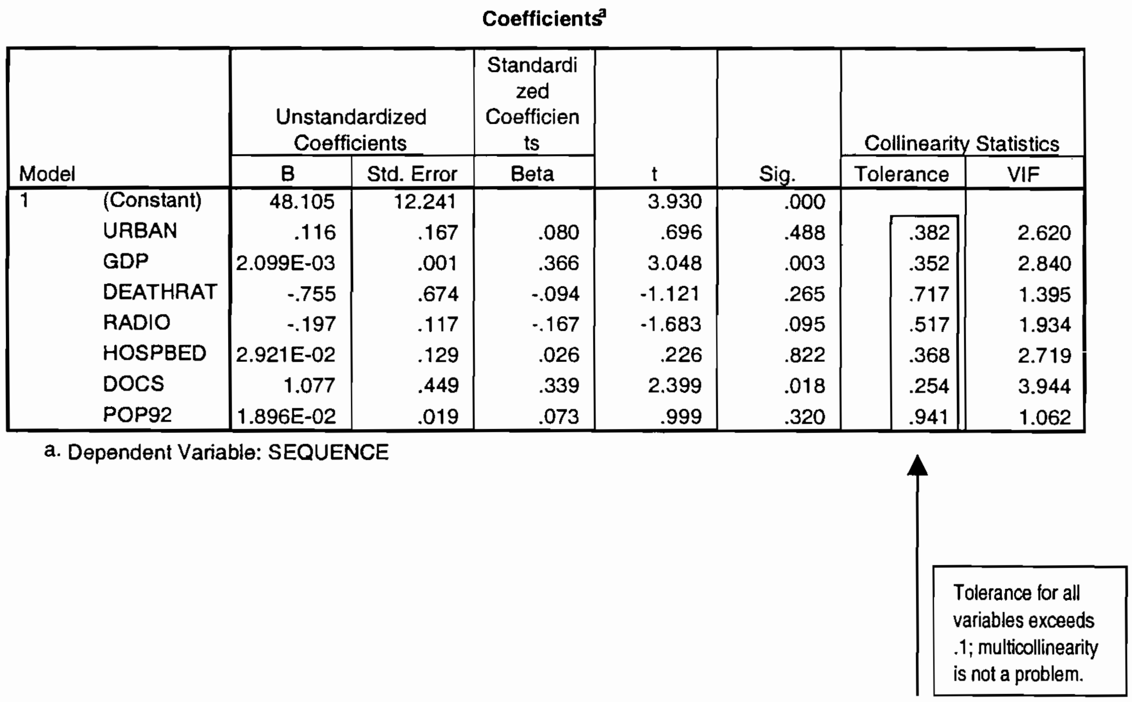
The output for logistic regression can be divided into three parts: the statistics for overall model fit, a classification table, and the summary of model variables. Although these components were introduced briefly in Section 11.1, a more detailed description will be presented here. The reader should note that the output for logistic regression looks considerably different from previous statistical methods since the output is presented in text and not in pivot tables. One should also keep in mind that the output can vary depending upon the stepping method utilized in the procedure. If a stepping method is applied, you have the option of presenting the output for each step or limiting the output to the last step. When output for each step is selected, the three components will be displayed for each step. Due to space constraints, our discussion of output and its subsequent interpretation will primarily be limited to output from the final step.

Several statistics for the overall model are presented in the first component of logistic regression output. The –2 Log Likelihood provides an index of model fit. A perfect model would have a –2 Log Likelihood of 0; consequently the lower this value, the better the model fits the data. This value actually represents the sum of the "probabilities associated with the predicted and actual outcomes for each case" (Tabachnik & Fidell, 1996, p. 582). The Goodness-of-Fit statistic is also calculated for the overall model and compares the predicted values of the subjects to their actual values. This value should also be relatively small. The next two values, Cox & Snell — R^2 and Nagelkerke — R^2, represent two different estimates of the amount of variance in the DV accounted for by the model. Chi-square statistics with levels of significance are also computed for the model, block, and step. Chi-square for the model represents the difference between the constant-only model and the model generated. When using a stepwise method, the model generated will include only selected predictors; in contrast, the enter method generates a model with all the IVs included. Consequently, this comparison varies depending on the stepping method utilized. In general, a significant model chi-square indicates that the generated model is significantly better in predicting subject membership than the constant-only model. However, the reader should note that a large sample size increases the likelihood of finding significance when a poor-fitting model may have been generated. Chi-square is also calculated for each step if a stepping method has been utilized. This value indicates the degree of model improvement when adding a selected predictor or, in other words, represents the comparison between the model generated from the previous step to the current step.

The second component of output to interpret is the classification table. This table applies the generated regression model to predicting group membership. These predictions are then compared to the actual subject values. The percent of subjects correctly classified is calculated and serves as another indicator of model fit.

The third component of output is the summary of model variables. This summary presents several statistics—*B, S.E., Wald, df, Sig., R, Exp(B)*—for each variable included in the model as well as the constant. *B*, as in multiple regression, represents the unstandardized regression coefficient and represents the effect the IV has on the DV. *S.E.* is the standard error of *B*. *Wald* is a measure of significance for *B* and represents the significance of each variable in its ability to contribute the model. Since several sources indicate that the *Wald* statistic is quite conservative (Tabachnick & Fidell, 1996), a more liberal significance level (i.e., $p<.05$ or $p<.1$) should be applied when interpreting this value. Degrees of freedom (*df*) and level of significance (*Sig.*) are also reported for the *Wald* statistic within the summary table. The partial correlation (*R*) of each IV with the DV (independent from the other model variables) is also presented. The final value presented in the summary table is *Exp(B),* which is the calculated odds ratio for each variable. The odds ratio represents the increase (or decrease if *Exp(B)* is less than 1) in odds of being classified in a category when the predictor variable increases by 1.

Figure 11.4 Tolerance Statistics for Example Number 1.



Coefficients[a]

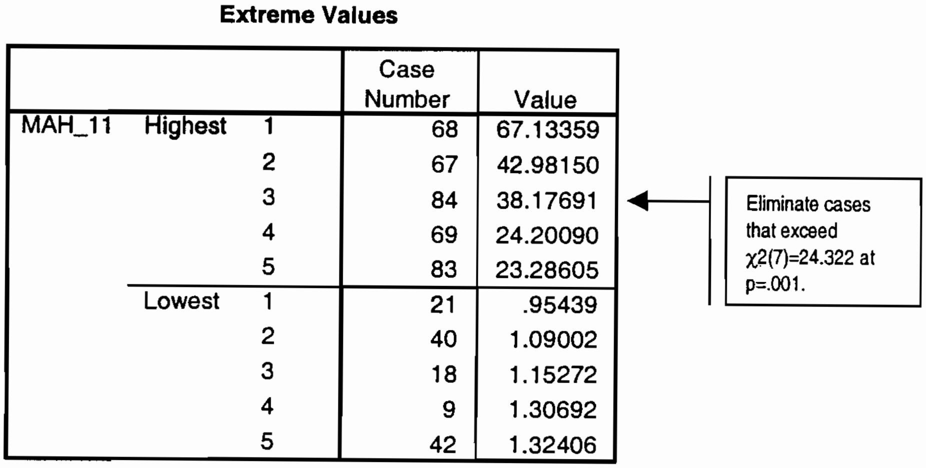
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	48.105	12.241		3.930	.000		
	URBAN	.116	.167	.080	.696	.488	.382	2.620
	GDP	2.099E-03	.001	.366	3.048	.003	.352	2.840
	DEATHRAT	-.755	.674	-.094	-1.121	.265	.717	1.395
	RADIO	-.197	.117	-.167	-1.683	.095	.517	1.934
	HOSPBED	2.921E-02	.129	.026	.226	.822	.368	2.719
	DOCS	1.077	.449	.339	2.399	.018	.254	3.944
	POP92	1.896E-02	.019	.073	.999	.320	.941	1.062

a. Dependent Variable: SEQUENCE

Tolerance for all variables exceeds .1; multicollinearity is not a problem.

Applying this process to our original example, we sought to investigate which IVs (population; percent of population living in urban areas; gross domestic product; death rate; number of radios, hospital beds, and doctors) are predictors of status as a developing nation (i.e., develop*ed* or develop*ing*). Since our investigation is exploratory in nature, we utilized the forward stepping method, such that only IVs that significantly predict the DV will be included in the model. Data were first screened for missing data and outliers. A preliminary multiple `Regression` was conducted to calculate Mahalanobis distance (to identify outliers) and examine multicollinearity among the seven predictors. Figure 11.4 presents the tolerance statistics for the seven predictors. Tolerance for all variables is greater than .1, indicating that multicollinearity is not a problem. The `Explore` procedure was then conducted to identify outliers (see Figure 11.5). Subjects with a Mahalanobis distance greater than $\chi^2(7)$=24.322 were eliminated. **Binary Logistic Regression** was then performed using the **Forward:LR** method. The three output components were presented in Figures 11.1 –11.3. Figure 11.1 indicates that the three variables, *gdp, hospbed,* and *urban* were entered into the overall model, which correctly classified 95.54% of the cases (see Figure 11.2). Figure 11.3 presents the summary of statistics for the model variables. Odds ratios, *Exp(B)* or e^B, indicated that as the variable urban increases by 1, subjects are 1.145 times more likely to be classified as "developing." The odds ratios for *gdp* and *hospbed* were both below 1, indicating that as *gdp* (e^B=.9990) and *hospbed* (e^B =.9202) increase by 1, the odds of being classified as "developing" decrease by the respective ratio.

Figure 11.5 Outliers for Mahalanobis Distance (Example Number 1).



Extreme Values

			Case Number	Value
MAH_11	Highest	1	68	67.13359
		2	67	42.98150
		3	84	38.17691
		4	69	24.20090
		5	83	23.28605
	Lowest	1	21	.95439
		2	40	1.09002
		3	18	1.15272
		4	9	1.30692
		5	42	1.32406

Eliminate cases that exceed $\chi2(7)$=24.322 at p=.001.

Writing Up Results

The results summary should always describe how variables have been transformed or deleted. The results for the overall model are reported within the narrative by first identifying the predictors entered into the model and addressing the following goodness of fit indices: -2 Log Likelihood, Goodness of Fit, and Model Chi-Square with degree of freedom and level of significance. The accuracy of classification should also be reported in the narrative. Finally, the regression coefficients for model variables should be presented in table and narrative format. The table should include *B, Wald, df,* level of signficance, and odds ratio. The following results statement applies the example presented in Figures 11.1 – 11.3.

> Forward logistic regression was conducted to determine which independent variables (population; percent of population living in urban areas; gross domestic product; death rate; number of radios, hospital beds, and doctors) were predictors of status as a developing nation (developed or developing). Data screening led to the elimination of three outliers. Regression results indicate the overall model of three predictors (gdp, hospital beds, and urban) was statistically reliable in distinguishing between developed and developing countries (-2 Log Likelihood=25.211; Goodness-of-Fit=26.407; $\chi^2(2)$=96.164, p<.0001). The model correctly classified 95.54% of the cases. Regression coefficients are presented in Table 1. Wald statistics indicated that all variables significantly predict country development. However, odds ratios for these variables indicate little change in the likelihood of country development.

Table 1. Regression Coefficients

	B	*Wald*	*df*	*p*	Odds Ratio
Urban	.1354	4.08	1	.0434	1.145
GDP	–.0010	6.60	1	.0102	.999
Hospital beds	–.0832	9.41	1	.0022	.920
Constant	3.0472	5.93	1	.0149	

SECTION 11.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example of the process of conducting logistic regression. This process includes the development of research questions, data screening methods, analysis methods, interpretation of output, and presentation of results. The example utilizes the data set *gss.sav* from the SPSS Web site.

Problem

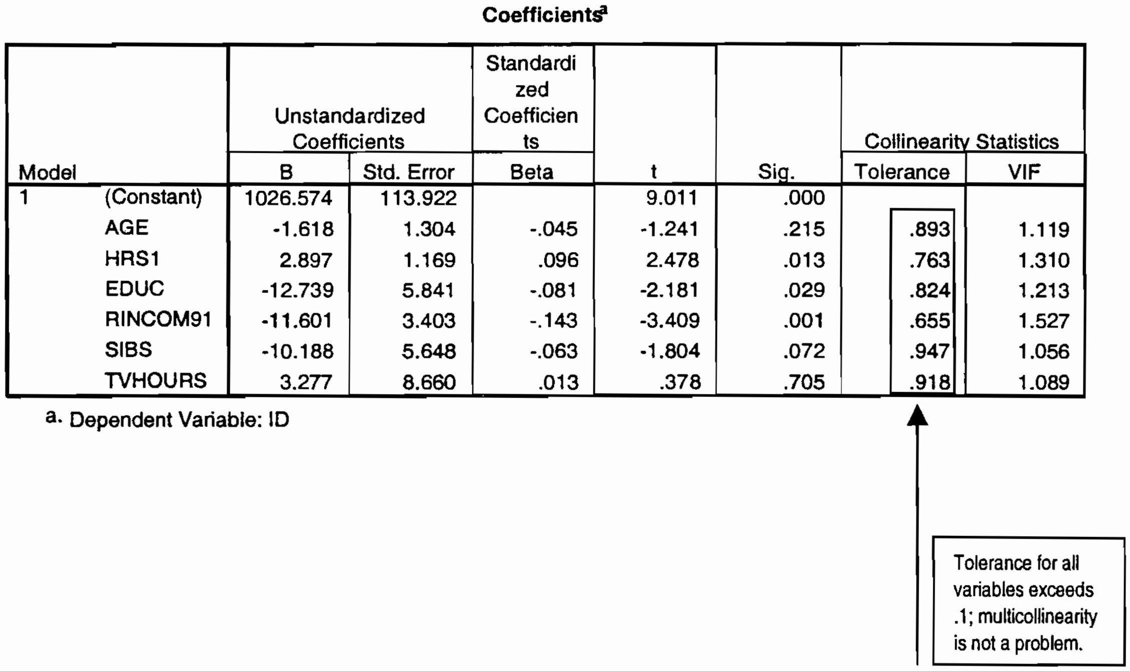
In the previous chapter on discriminant analysis, we presented the second example that investigated the ability of seven IVs (age, gender, hours worked per week, years of education, income, number of siblings, and number of hours spent watching TV) to predict one's life perspective (dull, routine, exciting). For this example, we will utilize a similar scenario (*excluding the gender variable*); however, the DV will be recoded as dichotomous to fulfill the requirement of binary logistic regression. Since six IVs are being investigated, the forward stepping method will be applied. The following research question is generated to address this scenario:

> Can *life* perspective (dull/routine or exciting) be reliably predicted from the knowledge of an individuals age (*age*), hours worked per week (*hrs1*), years of education (*educ*), income (*rincom91*), number of siblings (*sibs*), number of hours spent watching TV per day (*tvhours*)?

Method

Prior to analysis, the variable of *life* was recoded as dichotomous (*life2*) and applied the following transformations: 0=missing, 1-2=0, 3=1, 8-9=missing. Data were screened for missing data and outliers. A preliminary multiple **`Linear Regression`** was conducted to calculate Mahalanobis' Distance and to evaluate multicollinearity among the six continuous predictors. The table of regression coefficients (see Figure 11.6) indicate that multicollinearity was not violated since tolerance statistics for all six IVs were greater than .1. **`Explore`** was then conducted to determine which cases exceeded the chi square criteria of $\chi^2(6)=22.458$ at $p=.001$. All subjects that exceeded this value were eliminated from the analysis (see Figure 11.7). **`Binary Logistic Regression`** was then conducted using Forward: LR method.

Figure 11.6 Tolerance Statistics for Example Number 2.



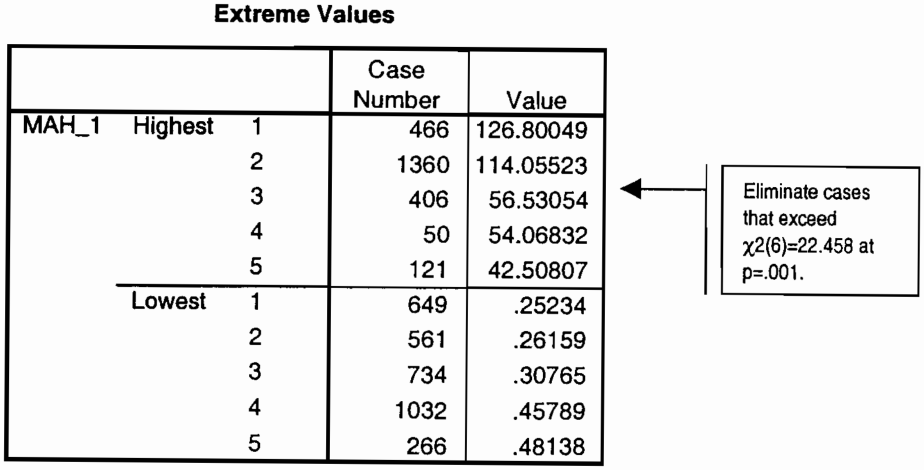
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1026.574	113.922		9.011	.000		
	AGE	-1.618	1.304	-.045	-1.241	.215	.893	1.119
	HRS1	2.897	1.169	.096	2.478	.013	.763	1.310
	EDUC	-12.739	5.841	-.081	-2.181	.029	.824	1.213
	RINCOM91	-11.601	3.403	-.143	-3.409	.001	.655	1.527
	SIBS	-10.188	5.648	-.063	-1.804	.072	.947	1.056
	TVHOURS	3.277	8.660	.013	.378	.705	.918	1.089

a. Dependent Variable: ID

Tolerance for all variables exceeds .1; multicollinearity is not a problem.

Figure 11.7 Outliers for Mahalanobis Distance (Example Number 2).



Extreme Values

			Case Number	Value
MAH_1	Highest	1	466	126.80049
		2	1360	114.05523
		3	406	56.53054
		4	50	54.06832
		5	121	42.50807
	Lowest	1	649	.25234
		2	561	.26159
		3	734	.30765
		4	1032	.45789
		5	266	.48138

Eliminate cases that exceed $\chi2(6)=22.458$ at p=.001.

Output and Interpretation of Results

The three components of output are presented in Figures 11.8–11.10. The statistics for overall model fit are presented in Figure 11.8 and indicate that only two variables were entered into the model: *educ* and *rincom91*. Model fit statistics were extremely large and reveal a poor-fitting model, -2Log Likelihood=748.595, Goodness of Fit=564.598. The generated model was significantly different from the constant-only model, $\chi^2(1)=33.098$, p<.0001. Figure 11.9 presents the classification table and indicates that the model correctly classified only 59.57% of subjects. The summary of model variables is displayed in Figure 11.10. Odds ratios for the *educ* (e^B=1.1609) and *rincom91* (e^B=1.0403) revealed little increase in the likelihood of perceiving life as exciting when the predictors increase by 1.

Figure 11.8 Goodness-of-fit Indices for Example Number 2.

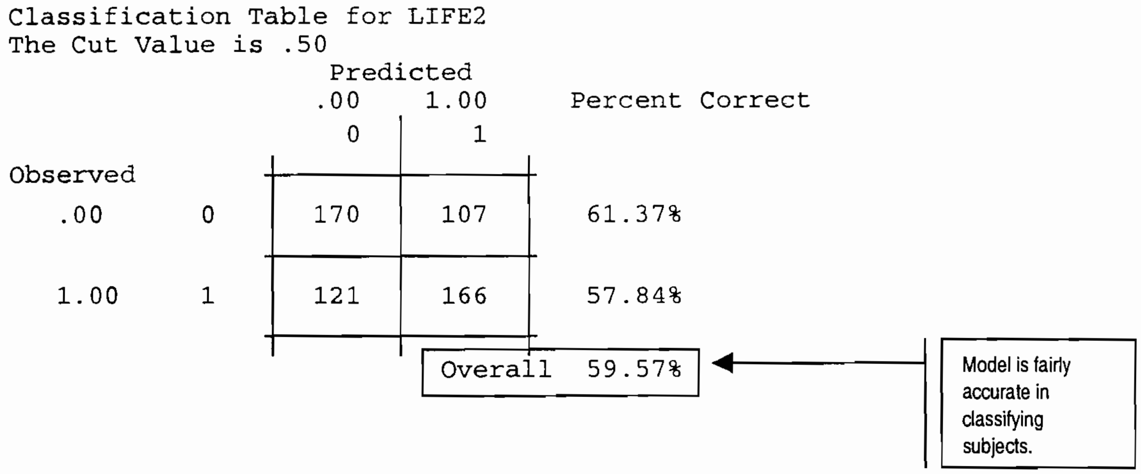
```
Dependent Variable..   LIFE2

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   781.69271

* Constant is included in the model.


Beginning Block Number  1.  Method: Forward Stepwise (LR)


       Improv.              Model           Correct
Step   Chi-Sq.  df    sig   Chi-Sq.  df    sig Class %    Variable
   1    27.898   1  .000     27.898   1  .000    59.04    IN: EDUC
   2     5.200   1  .023     33.098   2  .000    59.57    IN: RINCOM91

No more variables can be deleted or added.


End Block Number 1   PIN =     .0500  Limits reached.

Final Equation for Block 1

Estimation terminated at iteration number 3 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood          748.595
 Goodness of Fit            564.598
 Cox & Snell - R^2             .057
 Nagelkerke - R^2              .076

                     Chi-Square    df Significance

 Model                   33.098     1        .0000
 Block                   33.098     2        .0000
 Step                     5.200     1        .0226
```

Two variables were entered into the model.

Model fit indices are extremely large; fit is questionable.

Model significantly predicts group membership.

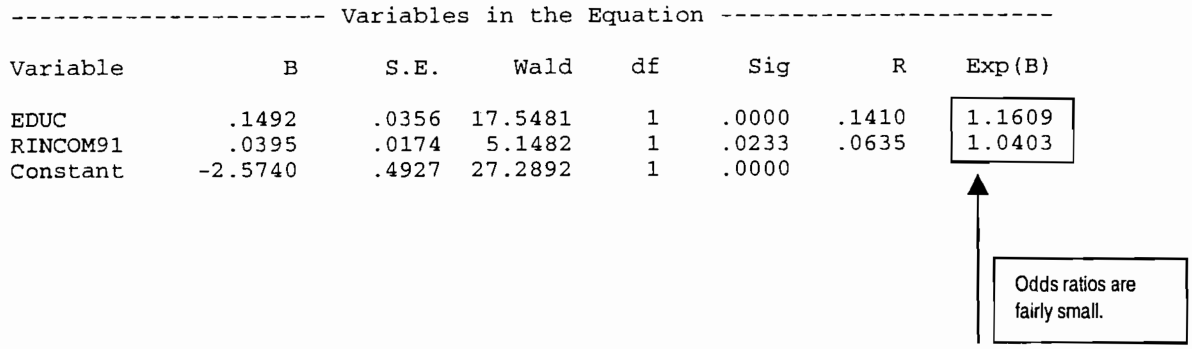
Figure 11.9 Classification Table for Example Number 2.



```
Classification Table for LIFE2
The Cut Value is .50
                                Predicted
                               .00     1.00      Percent Correct
                                0   |    1
Observed                     +------+-------+
   .00        0              | 170  |  107  |      61.37%
                             +------+-------+
   1.00       1              | 121  |  166  |      57.84%
                             +------+-------+
                                      Overall      59.57%
```

Model is fairly accurate in classifying subjects.

Figure 11.10 Regression Coefficients for Example Number 2.



```
---------------------- Variables in the Equation -----------------------

Variable           B        S.E.      Wald    df       Sig        R     Exp(B)

EDUC             .1492     .0356   17.5481     1     .0000    .1410     1.1609
RINCOM91         .0395     .0174    5.1482     1     .0233    .0635     1.0403
Constant       -2.5740     .4927   27.2892     1     .0000
```

Odds ratios are fairly small.

Presentation of Results

Forward logistic regression was conducted to determine which independent variables (age, hours worked per week, years of education, income, number of siblings, and number of hours spent watching TV) are predictors of life perspective (dull/routine or exciting). Data screening led to the elimination of several outliers. Regression results indicated the overall model fit of two predictors (education and income) was questionable (-2 Log Likelihood=748.595, Goodness of Fit=564.598) but was statistically reliable in distinguishing between life perspective; $\chi^2(1)=33.098$, p<.0001). The model correctly classified only 59.57% of the cases. Regression coefficients are presented in Table 1. *Wald* statistics indicated that education and income significantly predict life perspective. However, odds ratios for these variables indicated little change in the likelihood of perceiving life as exciting.

Table 1. Regression Coefficients

	B	*Wald*	*df*	*p*	Odds Ratio
Education	.1492	17.55	1	<.0001	1.1609
Income	.0395	5.15	1	.0233	1.0403
Constant	–2.5740	27.29	1	<.0001	

SECTION 11.5 SPSS "HOW TO"

This section demonstrates the steps for conducting binary logistic regression with Example Number 2 of this chapter. Prior to conducting binary logistic regression, be sure to dichotomize (0,1) your DV. To conduct **Binary Logistic Regression**, select the following menus:

Analyze
  Regression
    Binary Logistic

Logistic Regression Dialogue Box (see Figure 11.11)

Once in this dialogue box, identify the DV (*life2*) and move it to the Dependent box. Identify each IV and move each to the Covariates box. Next, select the desired regression method. SPSS provides seven different methods, five of which are described as follows:

Enter—Enters all the IVs at once into the model, regardless of significant contribution. This method is useful if you have previously tested the IVs and want all of them to be entered.

Forward: LR—One of the most common methods. Enters IVs, one at a time. The likelihood-ratio is used to determine variable selection.

Forward: Wald— Enters IVs, one at a time. The *Wald* statistic is used to determine variable selection.

Backward: LR—All IVs are entered at once, then variables are removed one at a time. The likelihood-ratio is used to determine variable removal.

Backward: Wald— All IVs are entered at once, then variables are removed one at a time. The *Wald* statistic is used to determine variable removal.

For our example, we selected **Forward: LR**. Next, click **Categorical**.

Logistic Regression: Define Categorical Variable Dialogue Box (see Figure 11.12)

By default in logistic regression, SPSS treats any numerical variable as continuous. Consequently, when an IV is categorical, you need to specify how SPSS should address it. Once in this dialogue box, identify any categorical variables and move them to Categorical Covariates box. Then under Change Contrast, select the method of contrast. SPSS provides several contrast methods. The Indicator method is the default and is the most common. Three of the contrasting methods are described as follows.

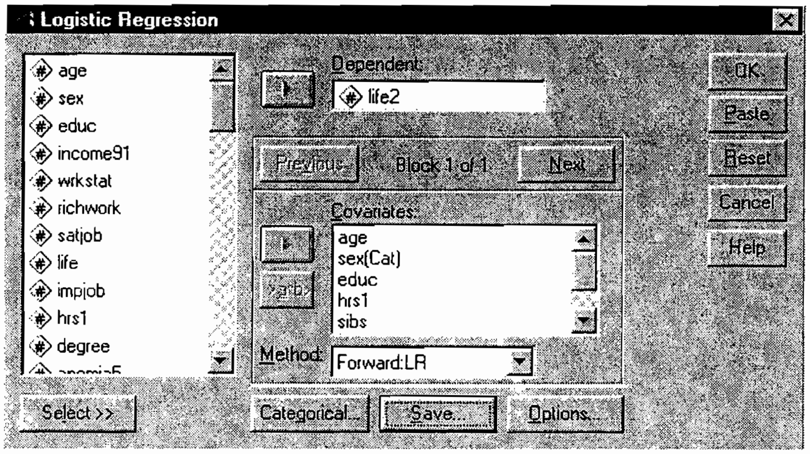
Indicator—Indicates the presence or absence of group membership. This is the default.

Simple—Each category of the IV (except the reference category) is compared to the reference category.

Deviation—Each category of the IV (except the reference category) is compared to the overall effect.

If you have selected one of these contrasting methods, you can identify a specific category to be used as the **Reference Category**. Two options are available: **Last** category (the default) or **First** category. If you have selected any options other than defaults for contrasting methods or reference category, you must then click **Change**. Click **Continue**, then **Options**.

Figure 11.11 Logistic Regression Dialogue Box.

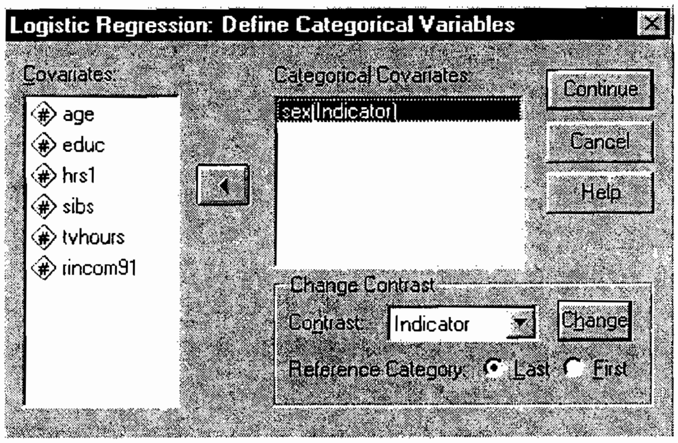


Dependent: life2

Covariates: age, sex(Cat), educ, hrs1, sibs

Method: Forward:LR

Figure 11.12 Logistic Regression: Define Categorical Variable Diaglogue Box.



Categorical Covariates: sex(Indicator)

Contrast: Indicator

Reference Category: ● Last ○ First

<u>**Logistic Regression: Options** Dialogue Box (see Figure 11.13)</u>

SPSS provides several options within logistic regression. Commonly used options are described below.

Classification Plots—Graph of actual and predicted values for the DV.

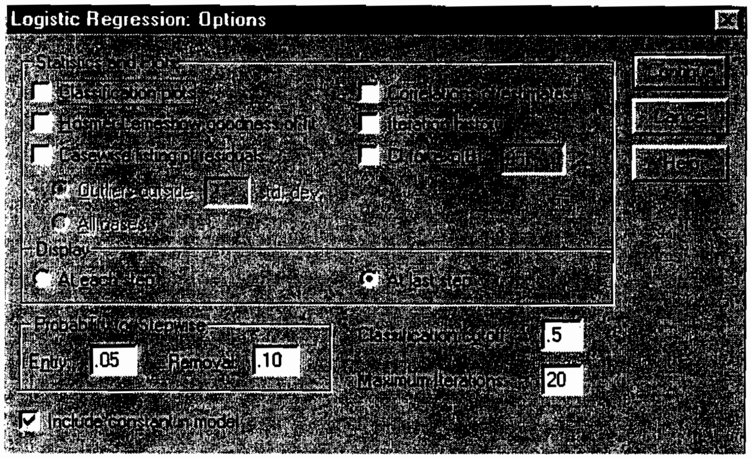
Correlations of Estimates—Correlation matrix of parameter estimates for model variables.

Iteration History—Presents coefficients and log likelihood at each iteration.

CI for exp(B)—Calculates the confidence intervals for the odds ratios of each model variable. You can indicate the level of probability associated with this interval. The default is 95%.

For our example, we did not select any of these options. A display option is also available if a stepping method has been utilized. You may want to select the display **At Last Step**, as we did, to conserve space. Other options available are the probability for stepwise, the maximum number of iterations, and inclusion of constant in the model. We maintained the defaults for these options; however, there may be times when it is necessary to increase the maximum number of iterations in order to generate a complete model. Once you have selected the appropriate options, click **Continue**, then **OK**. The reader should note that SPSS also provides options for saving variables. By clicking **Save**, you can save predicted values, residuals, etc., as new variables.

Figure 11.13 Logistic Regression: Options Dialogue Box.



Logistic Regression: Options

Statistics and Plots

□ Classification plots

□ Hosmer-Lemeshow goodness-of-fit

□ Casewise listing of residuals

○ Outliers outside std. dev.

○ All cases

□ Correlations of estimates

□ Iteration history

□ CI for exp(B)

Display

○ At each step

● At last step

Probability for Stepwise

Entry: .05 Removal: .10

Classification cutoff: .5

Maximum Iterations: 20

☑ Include constant in model

Summary

Logistic regression tests the ability of a model or group of variables to predict group membership as defined by some categorical DV. In binary logistic regression, the DV must be dichotomous, but the IVs may be categorical or continuous. Logistic regression actually predicts the probability of membership occurring, which varies from 0 to 1. A variety of methods can be used to test and develop different models (enter, Forward: LR, Backward: Wald, etc.). Although logistic regression requires fulfillment of few test assumptions, data should be screened for outliers and multicollinearity. Logistic regressions output includes three parts: statistics for overall model fit, classification table, and summary of model variables. Statistics for the overall model provide several indices of model fit: –2 Log Likelihood, Goodness-of-Fit, and Model Chi-Square. The classification table presents the percent of cases correctly classified with the generated model. The summary of model variables provides several variable statistics that indicate variable contribution to the model: *B, Wald, df,* level of significance, and odds ratio. A good fitting model will typically have: fairly low values for –2 Log Likelihood and Goodness-of-Fit, significant Model Chi-Square and variables with odds ratios greater than 1. Figure 11.14 provides a checklist for conducting binary logistic regression.

Figure 11.14 Checklist for Conducting Binary Logistic Regression.

I. Screen Data

a. Missing Data?

b. Multivariate Outliers and Multicollinearity?

❑ Run preliminary Linear Regression.

1. 🖱 **Analyze…Regression…Linear.**
2. Identify a variable that serves as a case number and move to Dependent Variable box.
3. Identify all appropriate quantitative variables and move to Independent(s) box.
4. 🖱 **Statistics.**
5. Check **Collinearity Diagnostics**.
6. 🖱 **Continue.**
7. 🖱 **Save.**
5. Check **Mahalanobis'.**
6. 🖱 **Continue.**
7. 🖱 **OK.**
8. Determine chi square χ^2 critical value at p<.001.

❑ Conduct **Explore** to test outliers for Mahalanobis chi square χ^2.

1. 🖱 **Analyze…Descriptive Statistics…Explore**
2. Move *mah_1* to Dependent Variable box.
3. Leave Factor box empty.
4. 🖱 **Statistics.**
5. Check **Outliers.**
6. 🖱 **Continue.**
7. 🖱 **OK.**

❑ Delete outliers for subjects when χ^2 exceeds critical χ^2 at p<.001.

II. Conduct Logistic Regression

a. Run Binary Logistic Regression using **Regression**.

1. 🖱 **Analyze…**🖱 **Regression…**🖱 **Binary Logistic…**
2. Move the DV to the Dependent Box.
3. Move IVs to the Covariates Box.
4. Select Method.
5. 🖱 **Categorical** (if any IVs are categorical).
6. Move any categorical IVs to the Categorical Covariates Box.
7. Select Contrast Method and Reference Category.
8. 🖱 **Continue.**
9. 🖱 **Options.**
10. Check appropriate options.
11. 🖱 **Continue.**
12. 🖱 **OK.**

III. Summarize Results

a. Describe any data elimination or transformation.

b. Describe the model generated (-2 Log Likelihood, Goodness of Fit, Model chi-square with *df* and *p*-value).

c. Report the accuracy of classification.

d. Present the regression coefficients for model variables in table format.

e. Report odds ratios for model variables.

f. Draw conclusions.

Exercises for Chapter 11

This exercise utilizes the data set *gss.sav*, which can be downloaded at the SPSS Web site. Open the URL: **www.spss.com/tech/DataSets.html** in your Web browser. Scroll down until you see "Data Used in SPSS Guide to Data Analysis—8.0 and 9.0" and click on the link "dataset.exe." When the "Save As" dialogue appears, select the appropriate folder and save the file. Preferably, this should be a folder created in the SPSS folder of your hard drive for this purpose. Once the file is saved, double-click the "dataset.exe" file to extract the data sets to the folder.

Conduct a Forward: LR logistic regression analysis with the following variables:

IV—*age, educ, hrsl, life, sibs, rincom91*
DV—*satjob2*

1. Develop a research question for the following scenario.

2. Conduct a preliminary `Linear Regression` to identify outliers and evaluate multicollinearity among the five continuous variables. Complete the following:

   a. Using the Chi-Square table, identify the critical value at p<.001 for identifying outliers. Use `Explore` to determine if there are outliers. Which cases should be eliminated?

   b. Is multicollinearity a problem among the five continuous variables?

3. Conduct `Binary Logistic Regression` using the Forward: LR method. Be sure to identify the variable of *life* as categorical (use the defaults).

   a. Which variables were entered into the model?

   b. To what degree does the model fit the data? Explain.

   c. Is the generated model significantly different from the constant-only model?

   d. How accurate is the model in predicting job satisfaction?

   e. What are the odds ratios for the model variables? Explain.