

Application of Logistic Regression in Health and Allied Research: Part II – Binary Logistic Regression

Mohammad Tajul Islam¹, Russell Kabir², Sufia Nasrin Rita³, Ali Davod Parsa², SM Anwar Sadat⁴

¹Department of Public Health, North South University, Dhaka, Bangladesh

²School of Allied Health and Social Care, Anglia Ruskin University, UK

³Department of Orthodontics, Sapporo Dental College, Dhaka, Bangladesh

⁴Department of Oral and Maxillofacial Surgery, Dhaka Dental College, Dhaka, Bangladesh

Corresponding author: Dr. Russell Kabir, School of Allied Health and Social Care, Anglia Ruskin University, Chelmsford, Essex, UK. Email: russell.kabir@aru.ac.uk

Abstract

Background: Binary logistic regression is a widely used multivariable method in health and allied research when the outcome is dichotomous. However, its appropriate application requires careful consideration of variable coding, model specification, assumptions, diagnostics, and interpretation. This article is Part II of a series on logistic regression and provides practical guidance for non-statisticians, clinical professionals, and early-career researchers.

Methods: This methodological tutorial describes the application of unconditional and conditional binary logistic regression using Stata. It covers coding of dichotomous outcomes and categorical predictors, selection of reference categories, model fitting, assessment of multicollinearity and model performance, estimation of predicted probabilities, variable selection, and interaction terms. Interpretation of regression coefficients, adjusted odds ratios, 95% confidence intervals, and post-estimation results is illustrated using practical examples.

Results: The tutorial demonstrates how binary logistic regression can be used to estimate adjusted associations and predict dichotomous outcomes. Practical examples illustrate the interpretation of adjusted odds ratios for categorical and continuous predictors, assessment of multicollinearity, evaluation of model calibration and discrimination, estimation of predicted probabilities, and examination of interaction terms. Important methodological considerations, including model diagnostics, overfitting, and variable selection, are highlighted.

Conclusions: Binary logistic regression is a flexible method for analyzing dichotomous outcomes in health and allied research. Appropriate model specification, diagnostics, and interpretation are essential for obtaining valid and meaningful results. This tutorial provides practical guidance for applying and reporting binary logistic regression using Stata.

Keywords: Binary logistic regression; Odds ratio; Stata; Multivariable analysis

1. Introduction

Part I of the logistic regression series discussed the basic concepts of the procedure, such as the purpose, methods, mathematical concept, variable requirements, sample size, and assumptions. In part II, we focus our discussion on how to perform the binary logistic regression using *Stata (version 17) software*, including the interpretation of outputs and use of post-estimation commands for verification of the model's effectiveness (logistic regression diagnostics).

Binary logistic regression analysis is appropriate when the outcome variable is a single *dichotomous categorical variable* (e.g., disease present/absent). For logistic regression analysis in Stata, the dichotomous outcome variable must be coded as 0/1 with "0= disease absent" and "1= disease present". Stata considers the higher value to be the predicted outcome and the lower value as the comparison group. Binary logistic regression can be performed as *unconditional and conditional* logistic regression methods (Islam, Kabir and Nisha, 2022). The unconditional logistic regression is used for unmatched studies, while the matched logistic regression is performed for matched designs, e.g., matched case-control studies.

2. Unconditional Binary Logistic Regression

Unconditional binary logistic regression is commonly expressed simply as "logistic regression". We will use the term "logistic regression" to denote unconditional binary logistic regression throughout this article.

Suppose that a researcher conducted an unmatched study (e.g., a cross-sectional or an unmatched case-control study) to identify the factors (or predictors) associated with asthma (dependent or

Table1. Codebook of variables in the data file

Variable name	Variable label	Variable codes
asthma	Has asthma (dependent variable)	0= no; 1= yes
age	Age in years	Actual value
sex	Sex	1= male; 2= female
religion	Religion of subjects	1= Muslim; 2= Hindu; 3= Others
fhistory	Has family history of asthma	0= no; 1= yes

outcome variable). The independent (also called explanatory or predictor) variables considered for the logistic regression analysis include age (as a continuous variable), sex (variable name: sex), religion (variable name: religion), and family history of asthma (variable name: fhistory). The variable names, their labels, and coding schemes are provided in Table 1. Data used for the demonstration is available in the data file <Asthma.dta> and can be downloaded from [Google Drive Link](#). To check the coding scheme of categorical variables, use any of the following commands:

```
label list asthma sex religion fhistory  
codebook asthma sex religion fhistory
```

2.1 Association between Sex and Asthma

Before performing the logistic regression analysis, let us see the association between sex and asthma by using the following commands:

```
tab sex asthma, row chi
tabodds asthma sex, base(2) or
```

The first command will generate a 2 by 2 table of asthma and sex with row percentages and the chi-square test results, while the second command will provide the odds ratio (OR) for males compared to females, including its 95% confidence interval (CI) and p-value (Table 2). In the second command, we used the option “base(2)” to indicate to Stata to use females as the comparison group (because females are coded as 2). If you don’t use this option, Stata will provide the OR for females compared to males (Stata considers the first category of a categorical variable as the comparison group by default).

Table 2 shows that the prevalence (consider the data are from a cross-sectional study) of asthma in males and females are 31.58% and 15.32%, respectively, and the difference is statistically significant, as indicated by the chi-square test p-value of 0.007. The OR for asthma for males is 2.55 (95% CI: 1.26 - 5.14; p=0.0067) compared to females (an OR of 1.00 indicates that females are the comparison group). This is the unadjusted or crude OR (cOR). We will compare it with the adjusted OR (aOR) given by the logistic regression analysis.

2.2 Performing the Analysis

Stata software (version 17) will be used to perform the logistic regression analysis using the data from the data file “Asthma.dta”. In the analysis, we will use asthma as the dependent (or outcome) variable. The independent (or explanatory) variables to be included are one continuous variable (age) and three categorical variables (sex, religion, and fhstory).

Table 2. Bivariate analysis for the association between sex and asthma

. tab sex asthma, row chi				
sex	has asthma			
	No	Yes	Total	
male	52	24	76	
	68.42	31.58	100.00	
female	105	19	124	
	84.68	15.32	100.00	
Total	157	43	200	
	78.50	21.50	100.00	
Pearson chi2(1) = 7.3781 Pr = 0.007				
. tabodds asthma sex, base(2) or				
sex	Odds ratio	chi2	P>chi2	[95% conf. interval]
male	2.550607	7.34	0.0067	1.263640 5.148301
female	1.000000	.	.	.
Test of homogeneity (equal odds): chi2(1) = 7.34				
Pr>chi2 = 0.0067				

Score test for trend of odds:	chi2(1)	=	7.34
	Pr>chi2	=	0.0067

To perform the analysis, recode the dependent variable (asthma) as 0/1 (0 for “no asthma” and 1 for “have asthma”) if it is not coded like this. For dichotomous independent variables, it is not mandatory to have the coding scheme 0/1. When the coding scheme of a variable is 0/1, Stata considers it a dichotomous variable. However, when a different coding scheme (e.g., 1/2) is used for a dichotomous variable, it is necessary to use the prefix “i.” for the variable to declare it a categorical variable. The prefix “i.” must also be used for multinomial independent variables for the analysis; otherwise, Stata will consider them continuous variables.

Logistic regression by Stata can be done using the commands “logit” and “logistic”. The first command (logit) provides the regression coefficients, which are difficult to interpret. The regression coefficients are used to calculate the predicted probabilities of the outcome with the independent variables in the analysis (model). The “logistic” command provides adjusted ORs, which are easier to interpret and is preferred. We should prefer using the first (logistic) of the following commands to perform the analysis (the second command will provide the regression coefficients, as shown in Table 5):

```
logistic asthma c.age i.sex i.religion i.fhhistory
logit asthma c.age i.sex i.religion i.fhhistory
```

The variable immediately after the command must be the dependent variable. In the above commands, we used prefixes “c.” and “i.” for the independent variables. The prefix “c.” is used to indicate that the variable is a continuous variable (e.g., age), while the prefix “i.” is used to indicate that the variable is a categorical variable (e.g., sex and others). The above commands will consider the first category (the lowest code) of all the categorical variables (by default) as the comparison group. For instance, the variable “sex” is coded as 1= male and 2= female. In this analysis, Stata will calculate and provide the OR for females compared to males. However, you can change the comparison group by using the prefix “ib.”. For instance, if you want females (coded as 2) as the comparison group, use the prefix “ib2.” for the variable “sex”, as shown in the following command:

```
logistic asthma c.age ib2.sex i.religion i.fhhistory
```

The outputs of the above command are provided in Table 3. Occasionally, in cross-sectional studies, data are collected using the cluster sampling method. In this situation, it is necessary to control (adjust) for the cluster effects. To adjust for cluster effects (say, the name of the cluster variable is “village”), use the following command:

```
logistic asthma c.age ib2.sex i.religion i.fhhistory, vce(cluster village)
```

2.3 Interpretation of Logistic Regression Outputs

In the analysis, we included four independent variables, of which one is a continuous variable (age) and three are categorical variables (sex, religion, and fhhistory). The outputs of the analysis are provided in Table 3.

Table 3 shows that data from 200 subjects are analyzed. The likelihood ratio (LR) chi-square [LR Chi2(5)] value is 96.24, and its p-value (Prob > chi2) is 0.000, which is statistically significant. A significant LR chi-square test indicates that the performance of the proposed model is significantly

better than the null model (i.e., a model without any independent variable) in predicting the outcome (Boateng and Abaye, 2019). On the contrary, if the LR chi-square test p-value were greater than 0.05, it would indicate that the independent variables are unable to predict the outcome, a situation where none of the independent variables in the model are statistically significant.

In linear regression, R^2 values indicate the proportion of variance in the dependent variable explained by the independent variables in the model, but pseudo R^2 values, which logistic regressions provide, cannot be interpreted like this. Pseudo R^2 values measure the model fit compared to the null model, like the likelihood ratio chi-square test. Pseudo R^2 values generally range from 0 to 1, with higher values indicating greater improvement in model fit relative to the null model. In this example, McFadden's pseudo R^2 was 0.462. However, this value should not be interpreted as the proportion of variance explained, as would be done with R^2 in linear regression.

Table 3. Logistic regression analysis with the dependent variable “asthma”

. logistic asthma c.age ib2.sex i.religion i.fhhistory						
Logistic regression			Number of obs = 200			
			LR chi2(5) = 96.24			
			Prob > chi2 = 0.0000			
Log likelihood = -55.980298			Pseudo R2 = 0.4623			
asthma	Odds ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.319234	.0595274	6.14	0.000	1.207573	1.44122
sex						
male	4.320893	2.274035	2.78	0.005	1.540267	12.12135
religion						
HINDU	1.956249	1.061979	1.24	0.216	.675051	5.669067
Christian	.5057676	.4934067	-0.70	0.485	.0747399	3.422546
fhhistory						
yes	3.067156	1.607687	2.14	0.033	1.097914	8.568472
_cons	8.38e-06	.0000145	-6.77	0.000	2.84e-07	.0002475
Note: _cons estimates baseline odds.						

However, reporting the pseudo R^2 value as a primary measure of model fit is often ignored by researchers, as it creates confusion with the R^2 value from linear regression. Other measures, such as the likelihood ratio test, classification table, and area under the ROC curve (see below), are used as a measure of model fit. The choice and emphasis of model diagnostics depend on the purpose of the analysis. Prediction models require particular attention to predictive performance, including discrimination and calibration, whereas models primarily intended to estimate adjusted associations still require appropriate assessment of model specification and assumptions.

Table 3 also shows the ORs for the independent variables with their standard errors (Std. Err.), p-values ($P>|z|$), and 95% CIs ([95% conf. interval]). *The ORs provided by the analysis are the adjusted ORs (aORs).* In our analysis, there are three categorical and one continuous independent variable. The interpretation of the aOR for categorical and continuous variables is different.

Let us first interpret the ORs for categorical variables. Table 3 shows that the OR for sex (being male) is 4.32 (95% CI: 1.54 – 12.12), which is statistically significant ($p= 0.005$). Here, the comparison group is female. Male participants had 4.32 times the odds of having asthma compared with female participants (aOR = 4.32; 95% CI: 1.54–12.12; $p = 0.005$), after adjusting for age, religion, and family

history of asthma. The crude (cOR) obtained from the bivariate analysis (Table 2) was 2.55, but after adjustment for confounding factors, it became 4.32. Note that the presence of one or more confounding factors can increase or decrease the effect size (Gordis, 2014 and Chan, 2004).

Individuals with a family history of asthma had 3.07 times the odds of having asthma compared with those without a family history (aOR = 3.07; 95% CI: 1.10–8.57; $p = 0.033$), after adjusting for the other variables in the model. The ORs for Hindus and Christians, compared to Muslims, are 1.95 (95% CI: 0.67 – 5.66; $p = 0.216$) and 0.505 (95% CI: 0.07 – 3.42; $p = 0.485$), respectively. The odds of asthma among Hindu participants were 1.96 times those among Muslim participants (aOR = 1.96; 95% CI: 0.68–5.67; $p = 0.216$); however, the association was not statistically significant. Similarly, Christians are 49.5% $[(1 - 0.505) \times 100]$ less likely to have asthma compared to Muslims after adjusting for age, sex, and family history of asthma, which is also not statistically significant.

The interpretation of the OR for age is different since the variable is entered as a continuous variable. In our example, the OR for age is 1.319 (95% CI: 1.207 – 1.441; $p = 0.000$). This indicates that the odds of having asthma increase (since the OR is greater than one) by 31.9% $[(1.319 - 1) \times 100]$; and the 95% CI is 20.7% - 44.1% with each year increase in age after controlling for all other variables (sex, religion, and family history of asthma) in the model, which is statistically significant ($p < 0.001$). If the OR for age were 0.80 (less than one), we would say that the odds of having asthma decrease by 20.0% $[(1 - 0.80) \times 100]$ with each year increase in age after controlling for other variables in the model.

2.4 Logistic Regression Diagnostics

In order to check whether or not the model is valid, it has to satisfy certain assumptions and requirements of logistic regression. When these assumptions and requirements are not met, the analysis may provide biased estimates of the coefficients (or the ORs), which can lead to misinterpretation of the results. Logistic regression diagnostics are a set of procedures for evaluating the validity and adequacy of a model by checking its assumptions, model fit, and other diagnostic measures.

2.4.1 Checking for multicollinearity

Multicollinearity occurs when one or more independent variables are in a linear combination (highly correlated) with other independent variables in the model. The degree of multicollinearity can vary and can have different effects on the model. When multicollinearity is present, the model becomes dubious, i.e., the model may not provide a correct estimate of regression coefficients or ORs.

The correlation coefficient (r value) is not a reliable measure for assessing multicollinearity in regression models. The standard and recommended method of checking multicollinearity among independent variables is to calculate the variance inflation factor (VIF) or its inverse ($1/\text{VIF}$), the tolerance value.

In Stata, there is no direct command to get these statistics after a logistic regression analysis. The VIF command in Stata is designed for use after linear regression. Therefore, to get the VIF (and the tolerance) statistic, it requires running a linear regression first using the same variables included in the logistic regression analysis. The dichotomous dependent variable (asthma) in this analysis is not a concern, as the purpose is to diagnose multicollinearity among independent variables. Use the following command to run the linear regression:

regress asthma c.age ib2.sex i.religion i.fhhistory

After performing the linear regression analysis, use the following command to get the VIF and tolerance statistics of the independent variables included in the analysis (Table 4):

vif

Table 4 shows that the tolerance (1/VIF) values for age, sex, religion (2: Hindu and 3: Christian), and family history of asthma are all greater than 0.88. Higher VIF values and lower tolerance values indicate greater multicollinearity among predictors. However, there is no universally accepted cut-off, and these statistics should be interpreted together with the study context and the relationships among predictors. In the present example, all VIF values are close to 1, providing no indication of problematic multicollinearity among the independent variables. When substantial multicollinearity is identified, researchers should examine the relationships among the affected predictors and consider appropriate approaches based on their substantive importance rather than automatically excluding variables solely on the basis of a numerical cut-off.

Another simple and subjective way to examine multicollinearity is to check the standard errors (SEs) of the coefficients as provided in Table 5. If multicollinearity is present and affects the model, the magnitude of standard errors of some of the coefficients will be very high (greater than 5.0) or very low (less than 0.001). In our example, there is no such problem. The existence of multicollinearity means that the model is not statistically stable. To solve the multicollinearity problem (in general),

Table 4. Multicollinearity among independent variables

. vif		
Variable	VIF	1/VIF
----- -----	-----	-----
age	1.05	0.951344
1.sex	1.09	0.917970
religion		
2	1.10	0.908336
3	1.11	0.903878
1.fhhistory	1.13	0.885400
----- -----	-----	-----
Mean VIF	1.10	

Table 5. Logistic regression coefficients with their standard errors

. logit asthma c.age ib2.sex i.religion i.fhhistory					
Iteration 0: log likelihood = -104.10128					
Iteration 1: log likelihood = -63.526382					
Iteration 2: log likelihood = -56.310705					
Iteration 3: log likelihood = -55.980955					
Iteration 4: log likelihood = -55.980298					
Iteration 5: log likelihood = -55.980298					
Logistic regression			Number of obs = 200		
			LR chi2(5) = 96.24		
			Prob > chi2 = 0.0000		
			Pseudo R2 = 0.4623		
Log likelihood = -55.980298					
asthma	Coefficient	Std. err.	z	P> z	[95% conf. interval]
----- -----	-----	-----	-----	-----	-----
age	.2770515	.0451227	6.14	0.000	.1886127 .3654903

sex							
male	1.463462	.5262883	2.78	0.005	.4319558	2.494968	
religion							
HINDU	.6710287	.5428649	1.24	0.216	-.392967	1.735024	
Christian	-.681678	.9755602	-0.70	0.485	-2.593741	1.230385	
fhistory							
yes	1.120751	.5241621	2.14	0.033	.0934118	2.148089	
_cons	-11.68978	1.727427	-6.77	0.000	-15.07548	-8.304088	

look at the SEs and omit the variable(s) with a very high (or very low) SE until the magnitude of the SEs hovers between 0.001 and 5.0 (Chan, 2004).

2.4.2 Checking for Model Fit

Model fit is a measure to indicate how close the predicted values are to the observed values. Logistic regression is commonly performed to adjust the ORs for confounding factors (adjustment model) or to identify predictors (prediction model) for the outcome variable. When the intention of the analysis is prediction, then the question is “How good is the model for prediction?” This is assessed based on the Hosmer-Lemeshow goodness-of-fit test and the positive and negative predictive values given by the classification table.

2.4.2.1 Hosmer-Lemeshow goodness-of-fit test

The Hosmer-Lemeshow test measures how well the observed and predicted values fit with each other (i.e., the observed and predicted values match with one another). The Hosmer-Lemeshow test evaluates whether there is evidence of lack of fit between observed and model-predicted outcomes across groups of predicted probabilities. A non-significant result indicates insufficient evidence of lack of fit based on this test. However, a non-significant result should not by itself be interpreted as proof of good predictive performance. Model performance should be evaluated using additional measures of calibration and discrimination. Use any of the following commands to get the statistic after performing the logistic regression analysis:

```

lfit, group(10)
estat gof, group(10)

```

The commands will generate Table 6. The Hosmer-Lemeshow test yielded a p-value of 0.316. This non-significant result indicates that there is insufficient evidence of lack of fit based on this test. Nevertheless, the Hosmer-Lemeshow test should not be used alone to determine whether a model has adequate predictive performance. **Table 6. Hosmer-Lemeshow goodness-of-fit test**

```

. lfit, group(10)
note: obs collapsed on 10 quantiles of estimated probabilities.

```

```

Goodness-of-fit test after logistic model
Variable: asthma

```

```

Number of observations =    200
Number of groups      =     10
Hosmer-Lemeshow chi2(8) =   9.32
Prob > chi2 = 0.3163

```

2.4.2.2 Classification table

The classification table provides the sensitivity, specificity, and positive and negative predictive values, including the overall accuracy of the model. The predictive values indicate how well the model is able to predict the correct category of the dependent variable (i.e., have or do not have the disease). To obtain the sensitivity, specificity, and positive and negative predictive values, we need to generate the classification table by using either of the following commands (after running the logistic regression analysis). Usually, the classification table is generated at a cut-off value of 0.5 (50%). You can, however, change the cut-off value of your choice.

lstat, cutoff(0.5)
estat classification, cutoff(0.5)

This command will generate Table 7. The table shows that the overall accuracy of this model to predict asthma (with a predicted probability of 0.5 or greater) is 89.0% (shown at the bottom of the table as “correctly classified”). The sensitivity and specificity of the model are 69.77% ($30 \div 43$) and 94.27% ($148 \div 157$), respectively, while the positive and negative predictive values are 76.92% ($30 \div 39$) and 91.93% ($148 \div 161$), respectively.

The measure of sensitivity indicates the model’s ability to correctly categorize individuals who have asthma as having a predicted probability value of >0.5 . A sensitivity of 69.77% means that among individuals who actually have asthma, the model will give a predicted probability value of >0.5 for 69.77% of them. Conversely, a specificity of 94.27% indicates that among individuals without asthma, the model will give a predicted probability value of <0.5 for 94.27% of them.

On the other hand, when the model gives a predicted probability value of >0.5 (the cutoff value) for an individual, the probability that the individual will actually have asthma is 76.92% (positive predictive value). On the contrary, when the model provides a predicted probability value of <0.5 for an individual, the probability that the individual will not have asthma is 91.93% (negative predictive value). For a detailed understanding of sensitivity, specificity, and positive and negative predictive values, readers may refer to any standard epidemiology book (Islam, Hawlader and Kabir, 2026; Gordis, 2014). You can also demonstrate the sensitivity and specificity at varying cut-off points in a graph by using the following command (outputs not shown):

lsens

Table 7. Classification table

. lstat, cutoff(0.5)				
Logistic model for asthma				
Classified	True			Total
	D	~D		
+	30	9		39
-	13	148		161
Total	43	157		200
Classified + if predicted Pr(D) >= .5				
True D defined as asthma != 0				
Sensitivity	Pr(+ D)			69.77%
Specificity	Pr(- ~D)			94.27%
Positive predictive value	Pr(D +)			76.92%
Negative predictive value	Pr(~D -)			91.93%

False + rate for true ~D	Pr(+ ~D)	5.73%
False - rate for true D	Pr(- D)	30.23%
False + rate for classified +	Pr(~D +)	23.08%
False - rate for classified -	Pr(D -)	8.07%
Correctly classified		89.00%

You can see (Table 7) that Stata did not provide the 95% CIs for sensitivity, specificity, and positive and negative predictive values in the output. To obtain the CIs, we need to install the “diagt” module with the following command:

```
ssc install diagt, replace
```

After installing the module, use the following commands to obtain the CIs:

```
predict pre_prob  
gen pre_outcome = pre_prob > 0.5  
diagt asthma pre_outcome
```

The first command will generate a new variable “pre_prob” with the values of predicted probabilities for individuals. The second command will generate a new variable “pre_outcome”, with the codes 0 (for predictive probability of ≤ 0.5) and 1 (for predictive probability of > 0.5). The third command will provide the classification table (a 2 by 2 table) with the sensitivity, specificity, and positive and negative predictive values, including their 95% CIs (outputs not shown).

The classification table is particularly useful when the purpose of the logistic regression model is prediction or classification. Its interpretation depends on the selected probability threshold and should be considered together with other measures of predictive performance..

2.4.2.3 ROC curve

We can construct the receiver operating characteristic (ROC) curve to assess the model discrimination, which is an indication of the accuracy of the logistic regression model in predicting the outcome with the independent variables. Discrimination is the ability of the model to distinguish between those who have the outcome (e.g., a disease) and those who do not have the outcome. Discrimination is evaluated by using the ROC curve analysis. In ROC curve analysis, the area under the curve (also called “C-statistic”) is measured. The AUC provides an overall measure of the model’s ability to discriminate between individuals with and without the outcome (Boateng and Abaye, 2019; Chan 2004).

The area under the ROC curve (AUC) ranges from 0 to 1. A value of 0.5 indicates the model is useless. Larger AUC indicates better predictability of the model. Values between 0.70 and 0.80 are considered acceptable discrimination, values between > 0.80 and 0.90 indicate excellent discrimination, and values > 0.90 indicate outstanding discrimination (Islam, Hawalder and Kabir, 2026). To obtain the ROC curve, after performing the logistic regression analysis, use the following command (Figure 1):

```
lroc
```

Figure 1, generated by the command, shows that the area under the curve is 0.914 (at the bottom of the figure), indicating an outstanding model for prediction.

We can also calculate the AUC with its 95% CI separately. To calculate the AUC, we need to generate a classifier variable first (say, class). Then we can calculate the AUC with its 95% CI. Use the following commands:

```
predict class
roctab asthma class
```

The first command will generate the classifier variable “class”, while the second command will calculate the AUC with its 95% CI (Table 8).

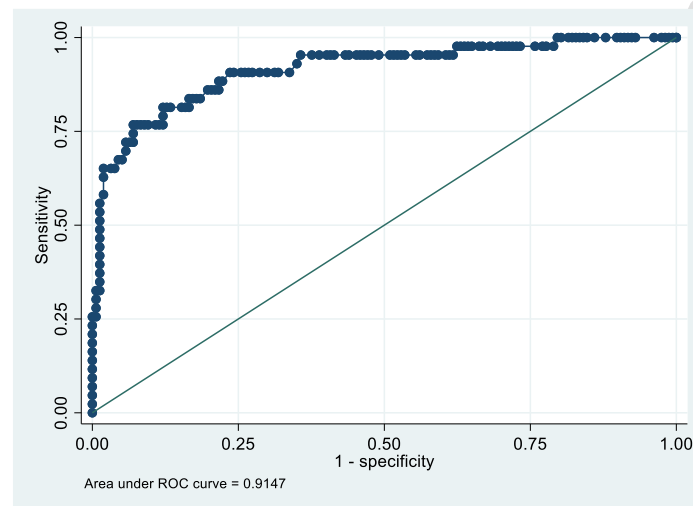


Figure 1. ROC curve for model discrimination

Table 8. Area under the ROC curve

. roctab asthma class				
Obs	ROC area	Std. err.	Asymptotic normal [95% conf. interval]	
200	0.9147	0.0269	0.86191	0.96744

2.5 Estimation of predicted values

Based on the logistic regression model, Stata can provide the estimated predicted probability of the outcome for an individual using the post-estimation commands. Post-estimation commands are those used after performing the primary analysis. Use the following command to obtain the estimated predicted probability for each individual in the dataset:

```
predict pre1, pr
```

This command will generate a new variable “pre1” with the predicted probability of having the outcome (in our example, asthma) for each individual in the dataset. The command calculates the probability based on the individual’s characteristics included in the model.

Stata can also provide the adjusted probability of having the outcome for independent variables in the model, considering the average values (sample means) of other variables. Use the following command

to obtain the predicted probability (adjusted probability) of having asthma for each category of religion, considering (adjusting for) the average values of other variables in the model (e.g., age, sex, and family history of asthma):

```
margins religion, atmeans
marginsplot
```

The first command will provide the predicted probability of asthma for each level of religion (Table 9). The table shows that the predicted probability (Margin) of asthma (i.e., the probability of having asthma) for being a Muslim is 0.071, while it is 0.130 for Hindus and 0.037 for Christians. These predicted values are calculated considering the average values (sample means, as shown in the table) of other variables (such as age, sex, and family history of asthma) in the model. The second command will display a graph of predicted probability for all categories of religion with their 95% CIs (Figure 2).

Table 9. Predicted probabilities for different categories of religion

. margins religion, atmeans

Adjusted predictions

Number of obs = 200

Model VCE: OIM

Expression: Pr(asthma), predict()

At: age = 29.03 (mean)

1.sex = .38 (mean)

2.sex = .62 (mean)

1.religion = .59 (mean)

2.religion = .28 (mean)

3.religion = .13 (mean)

0.fhistory = .535 (mean)

1.fhistory = .465 (mean)

	Margin	Delta-method std. err.	z	P> z	[95% conf. interval]	
religion						
MUSLIM	.0711134	.0280008	2.54	0.011	.0162327	.125994
HINDU	.1302576	.0543171	2.40	0.016	.0237981	.2367172
Christian	.037277	.0356693	1.05	0.296	-.0326336	.1071876

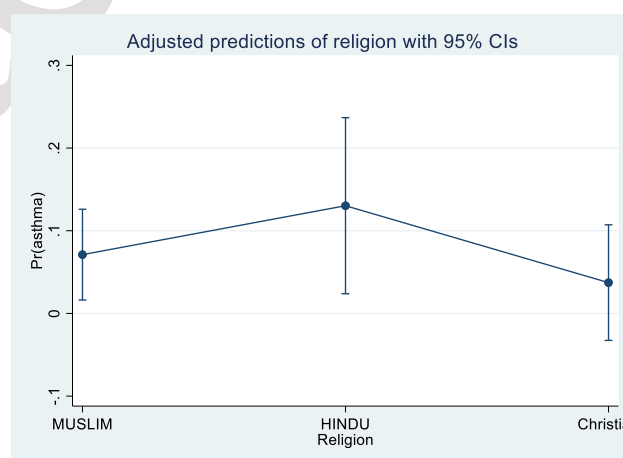


Figure 2. Predicted probabilities for religion with the 95% CIs

Other postestimation commands for obtaining the adjusted predicted probability for a variable, fixing one or more variables specified in the command, and holding other variables at their observed values, include:

```
margins religion, at(sex=(1 2))
```

```
margins sex, at(religion=(1(1)3))
```

```
margins sex, at(age=(30))
```

For example, the first command will provide the predicted probability for each category of religion for males and females, considering other variables in the model at their observed values (if you want them at mean values of other variables, use the option “, atmeans”). The following commands are for obtaining the risk difference, holding other variables at their observed values.

```
margins, dydx(i.sex c.age)
```

```
margins r.sex
```

2.6 Variable selection for a model

Like other multivariable analyses (e.g., multiple linear regression), independent variables to be selected for a logistic regression should include the risk factors of interest, potential confounders, and clinically relevant variables while avoiding variables with many missing values (Chan, 2004; Domínguez-Almendros et al, 2011).

Variable selection should primarily be guided by the research question, subject-matter knowledge, prior evidence, and the purpose of the model. Variables that are clinically or scientifically important, including prespecified potential confounders, should not be excluded solely because they are not statistically significant in univariable analyses. The number of parameters that can be reliably estimated also depends on the amount of information available in the data, particularly the number of outcome events. Therefore, variable selection should consider both substantive relevance and the risk of model overfitting.

The command that we used for the logistic regression is the “Enter” method, i.e., a method that uses all the independent variables in the model specified by the investigator (full model). We can also use the automatic selection method (stepwise method) for the analysis.

There are several automatic variable selection methods in regression modeling, of which the following three are the most common: (a) *Forward stepwise method*: this method starts with no independent variables in the model and then adds variables one by one according to a pre-established criterion (forward procedure); (b) *Backward stepwise method*: this method starts with all candidate variables in the model and then eliminates variables according to a pre-established criterion (backward procedure); and (c) *Stepwise method*: This method is a combination of above two approaches. At each step, it either adds a new variable to the model or removes an existing one, depending on the predefined criteria for entry or removal (Domínguez-Almendros et al, 2011).

For logistic regression, the commonly used method for the automatic selection of variables for modeling is the “*Backward Likelihood Ratio or Backward LR*” method, which uses the likelihood ratio test for variable removal. The other method of automatic selection of variables, particularly when the number of predictors is large, is the “*Forward LR*” method. However, some researchers criticize the use of stepwise methods because they can provide biased estimates and inflated Type I

errors (Malek et al, 2007; Whittingham et al, 2006). The commands for stepwise logistic regression include:

```
sw, pr(.1) lr: logistic asthma c.age ib2.sex i.religion i.fhistory
sw, pr(.1) pe(.05) forward lr: logistic asthma c.age ib2.sex i.religion i.fhistory
sw, pr(.1) lr lockterm1: logistic asthma (c.age i.religion) ib2.sex i.fhistory
sw, pr(.1) pe(.05) forward lr lockterm1: logistic asthma (c.age i.religion) ib2.sex i.fhistory
```

Using the likelihood ratio test (the preferred method), the first command is for backward LR, while the second command is for the forward LR method. The third and fourth commands are for forced entry of the variables “age” and “religion” into the model for backward LR and forward LR methods, respectively. The “pr(0.1)” indicates the removal criteria of independent variables from the model (i.e., if the p-value of a variable is >0.10, the variable will be removed from the model), while the “pe(0.05)” indicates the inclusion (entry) criteria of variables into the model (i.e., if the p-value of a variable is <0.05, the variable will be added to the model). You can change the criteria for removal and inclusion (entry) based on your needs. The interpretation of the outputs is the same as discussed earlier.

2.7 Incorporating interaction terms in the model

Interaction and confounding are not the same. Interaction is also called “effect modification”. Interaction is said to be present when the strength of association [OR or relative risk (RR)] of an independent variable for an outcome variable is different at different levels (categories) of a third variable. If the strength of association is the same at different levels of the third variable, there is no interaction. For details, readers are referred to (Islam, Hawlader and Kabir , 2026; Gordis, 2014 and Katz, 2011).

To assess interaction, we can include interaction terms in the logistic regression analysis with other variables for adjustment. Suppose that we are interested in assessing whether there is an association of sex and family history of asthma with asthma. We are also interested to know whether or not there is an interaction between sex and family history of asthma on the outcome. In this situation, we need to include both the independent variables (sex and family history of asthma) and their interaction terms in the model. Note that the variables for which we are looking for interaction must also be included in

Table 10. Logistic regression with interaction terms

. logistic asthma c.age ib2.sex i.religion i.fhistory ib2.sex#i.fhistory						
Logistic regression			Number of obs = 200			
			LR chi2(6) = 97.59			
			Prob > chi2 = 0.0000			
Log likelihood = -55.30662			Pseudo R2 = 0.4687			
asthma	Odds ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.316658	.0599243	6.04	0.000	1.204295	1.439505
sex						
male	2.375892	1.718608	1.20	0.232	.5755829	9.807211
religion						
HINDU	2.020169	1.114052	1.28	0.202	.685457	5.953813
Christian	.5353948	.5219784	-0.64	0.522	.0792141	3.618641
fhistory						

yes	1.799196	1.223937	0.86	0.388	.4742774	6.825342
sex#fhistory						
male#yes	3.303581	3.406356	1.16	0.246	.4378267	24.92686
_cons	.0000127	.0000223	-6.42	0.000	4.05e-07	.0003963

Note: _cons estimates baseline odds.

the model independently. To add the interaction term (sex#fhistory) into the model, use the following command:

logistic asthma c.age ib2.sex i.religion i.fhistory ib2.sex#i.fhistory

The above command will provide Table 10. The interaction term included in the model is indicated as “male#yes”. The table shows that the p-value for the interaction is 0.246 (>0.05), indicating that there is no significant interaction between sex and family history of asthma for the outcome, i.e., the effect of sex on asthma does not differ (not dependent on) by family history status, and vice versa. If there were an interaction (interaction p-value is <0.05), data would need to be analyzed separately at each level of sex or family history of asthma, depending on the objective. Note that when the interaction term is non-significant, the model should be rerun without the interaction term to obtain the main effects (aORs) of the independent variables.

You can obtain the predicted probabilities (estimated probabilities) for combinations of sex and family history of asthma (e.g., estimated probability of having asthma in a male with a positive family history of asthma, as well as other combinations) using the following command (Table 11):

margins ib2.sex#i.fhistory

Table 11 shows the predicted probabilities of asthma for combinations of sex and family history. Among males, the predicted probability of asthma was 0.416 (41.6%) among those with a family history of asthma and 0.216 (21.6%) among those without a family history. Among females, the corresponding predicted probabilities were 0.189 (18.9%) and 0.140 (14.0%), respectively. Although these predicted probabilities differ across combinations of sex and family history, the interaction term was not statistically significant ($p = 0.246$). Therefore, these data do not provide sufficient statistical evidence of effect modification on the modeled scale.

Table 11. Predicted probabilities for combinations of sex and family history of asthma

. margins ib2.sex#i.fhistory						
Predictive margins				Number of obs = 200		
Model VCE: OIM						
Expression: Pr(asthma), predict()						

		Margin	Delta-method std. err.	z	P> z	[95% conf. interval]
-----+-----						
sex#fhistory						
male#no		.2156268	.0462701	4.66	0.000	.1249391 .3063144
male#yes		.4161626	.0807076	5.16	0.000	.2579786 .5743466
female#no		.1400302	.0420626	3.33	0.001	.057589 .2224715
female#yes		.1894983	.036197	5.24	0.000	.1185534 .2604432

3. Conditional binary logistic regression

The conditional logistic regression technique is employed when cases and controls are matched for one or more variables in a matched case-control study. Suppose that we conducted a matched case-control study to identify the risk factors for death due to acute myocardial infarction (AMI). In this study, cases are matched with controls by sex. The outcome of interest is death due to AMI (case), and controls are those who survived an attack of AMI. Each case (case-to-control ratio is 1:1) is matched with a control based on sex. The risk factors to be evaluated in this study are age, religion, diabetes, and hypertension. The data for this matched case-control study is given in the data file **<Data matched.dta>**.

In the dataset, the variable “mID” represents the matching ID number of cases and controls (we must have this variable). The variable “death” is categorized as cases (who died; coded as 1) and controls (who survived; coded as 0). The codebook of variables is provided in Table 12.

In the analysis, we will use the variables death (outcome or dependent variable), age (as a continuous variable), religion, diabetes, and hypertension (variable name: htn). *The variable “mID” is needed to specify the groups.* Since cases and controls are matched for sex, it is meaningless to include this variable (sex) in the analysis. To perform the conditional logistic regression, use the following command:

clogit death c.age ib3.religion i.diabetes i.htn, group(mID) or

The variable immediately after the command (clogit) must be the outcome (dependent) variable. The prefix “ib3.” used for religion, indicates that Christians (the 3rd category of religion) are the comparison group. The above command will generate Table 13 showing the adjusted ORs with their 95% CIs and p-values.

The interpretation of outputs is similar to unconditional logistic regression analysis, as discussed earlier. For example, the OR for Muslims is 13.0 (95% CI: 1.49-113.29; p=0.020), indicating that Muslim participants had 13.0 times the odds of death following AMI compared with Christian participants (aOR = 13.0; 95% CI: 1.49–113.29; p = 0.020), after controlling for age, diabetes, and hypertension. The wide 95% CI indicates considerable imprecision in the estimate, which may reflect limited information or sparse data for this comparison. Note that since cases and controls are matched for sex, the variable is automatically adjusted.

Table 12. Codebook for the data file “Data matched.dta”

Variable name	Variable label	Variable codes
id	ID number of subjects	Actual value
mID	Matching ID	Actual value
death	Death due to AMI	0= control/alive; 1= case/died
sex	Sex	0= female; 1= male
age	Age in years	Actual value
religion	Religion	1= Muslim; 2= Hindu; 3= Christian
diabetes	Have diabetes	0= no; 1= yes
htn	Have hypertension	0= no; 1= yes

Table 13. Conditional logistic regression with coefficients

. clogit death c.age ib3.religion i.diabetes i.htn, group(mID) or						
Iteration 0: log likelihood = -47.461808						
Iteration 1: log likelihood = -45.703307						
Iteration 2: log likelihood = -45.60376						
Iteration 3: log likelihood = -45.603507						
Iteration 4: log likelihood = -45.603507						
Conditional (fixed-effects) logistic regression				Number of obs = 200		
				LR chi2(5) = 47.42		
				Prob > chi2 = 0.0000		
Log likelihood = -45.603507				Pseudo R2 = 0.3421		
death	Odds ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.076955	.0278228	2.87	0.004	1.023781	1.132891
religion						
MUSLIM	13.0085	14.36527	2.32	0.020	1.493641	113.2943
HINDU	6.840675	7.467477	1.76	0.078	.8051986	58.1159
diabetes						
Yes	2.183024	.9155918	1.86	0.063	.959516	4.966664
htn						
Yes	2.569417	1.160048	2.09	0.037	1.060552	6.224974

You can check multicollinearity (using the tolerance value or VIF) among independent variables using the same techniques as discussed in Section 2.4.1 and use the pseudo R^2 value to assess the model fit (see Section 2.3). Hosmer-Lemeshow goodness-of-fit test and classification table are not valid for conditional logistic regression (Chan, 2005). There is no direct method (i.e., a single command after the primary analysis) to obtain the AUC for the matched design. Stepwise methods can be used for automatic selection of variables by using the commands as discussed in Section 2.6. For instance, for the backward LR method, use the following command:

sw, pr(.05) lr: clogit death c.age ib3.religion i.diabetes i.htn, group(mID) or

The multinomial and ordinal regression methods will be discussed in the next issue (Application of Logistic Regression in Health and Allied Research: Part III – Multinomial and Ordinal regressions) of the publication.

References

- Boateng, E.Y. and Abaye, D.A. (2019) 'A review of the logistic regression model with emphasis on medical research', *Journal of Data Analysis and Information Processing*, 7, pp. 190–207. doi: 10.4236/jdaip.2019.74012.
- Chan, Y.H. (2004) 'Biostatistics 202: logistic regression analysis', *Singapore Medical Journal*, 45(4), pp. 149–153.
- Chan, Y.H. (2005) 'Biostatistics 305: multinomial logistic regression', *Singapore Medical Journal*, 46(6), p. 259.

- Domínguez-Almendros, S., Benítez-Parejo, N. and Gonzalez-Ramirez, A.R. (2011) 'Logistic regression models', *Allergologia et Immunopathologia*, 39(5), pp. 295–305. doi: 10.1016/j.aller.2011.05.002.
- Gordis, L. (2014) *Epidemiology*. 5th edn. Philadelphia: Elsevier Saunders.
- Islam, M.T., Hawlader, M.D.H. and Kabir, R. (2026) *Basic principles of epidemiology*. 2nd edn. New York: CRC Press.
- Islam, M.T., Kabir, R. and Nisha, M. (2022) *Data analysis with Stata: a comprehensive guide for data analysis and interpretation of outputs*. Dhaka, Bangladesh: Altaf Publications.
- Katz, M.H. (2011) *Multivariable analysis: a practical guide for clinicians and public health researchers*. 3rd edn. Cambridge: Cambridge University Press.
- Malek, M.H., Berger, D.E. and Coburn, J.W. (2007) 'On the inappropriateness of stepwise regression analysis for model building and testing', *European Journal of Applied Physiology*, 101(3), pp. 263–264. doi: 10.1007/s00421-007-0485-9.
- Whittingham, M., Stephens, P., Bradbury, R. and Freckleton, R. (2006) 'Why do we still use stepwise modelling in ecology and behavior?', *Journal of Animal Ecology*, 75(5), pp. 1182–1189.