Chapter 6
Logistic Regression: Regression with a Binary Dependent Variable

Chapter 6
Logistic Regression: Regression with a Binary Dependent Variable

LEARNING OBJECTIVES

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

- State the circumstances under which logistic regression should be used instead of multiple regression.
- Identify the types of dependent and independent variables used in the application of logistic regression.
- Describe the method used to transform binary measures into the likelihood and probability measures used in logistic regression.

Chapter 6
Logistic Regression: Regression with a Binary Dependent Variable

LEARNING OBJECTIVES continued . . .

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

- Interpret the results of a logistic regression analysis and assessing predictive accuracy, with comparisons to both multiple regression and discriminant analysis.
- Understand the strengths and weaknesses of logistic regression compared to discriminant analysis and multiple regression.

Logistic Regression Defined

Logistic Regression . . . is a specialized form of regression that is designed to predict and explain a binary (two-group) categorical variable rather than a metric dependent measure. Its variate is similar to regular regression and made up of metric independent variables. It is less affected than discriminant analysis when the basic assumptions, particularly normality of the independent variables, are not met.

Logistic Regression May Be Preferred . . .

When the dependent variable has only two groups, logistic regression may be preferred for two reasons:

- Discriminant analysis assumes multivariate normality and equal variance-covariance matrices across groups, and these assumptions are often not met. Logistic regression does not face these strict assumptions and is much more robust when these assumptions are not met, making its application appropriate in many situations.
- Even if the assumptions are met, some researchers prefer logistic regression because it is similar to multiple regression. It has straightforward statistical tests, similar approaches to incorporating metric and nonmetric variables and nonlinear effects, and a wide range of diagnostics.

Multiple Regression Decision Process

Stage 1: Objectives of Logistic Regression

Stage 2: Research Design for Logistic Regression

Stage 3: Assumptions of Logistic Regression

Stage 4: Estimation of the Logistic Regression Model and Assessing Overall Fit

Stage 5: Interpretation of the Results

Stage 6: Validation of the Results

Stage 1: Objectives of Logistic Regression

Logistic regression is best suited to address two research objectives . . .

- Identifying the independent variables that impact group membership in the dependent variable.

- Establishing a classification system based on the logistic model for determining group membership.

Stage 2: Research Design for Logistic Regression

- The binary nature of the dependent variable (0 – 1) means the error term has a binomial distribution instead of a normal distribution, and it thus invalidates all testing based on the assumption of normality.
- The variance of the dichotomous variable is not constant, creating instances of heteroscedasticity as well.
- Neither of the above violations can be remedied through transformations of the dependent or independent variables. Logistic regression was developed to specifically deal with these issues.

Stage 3: Assumptions of Logistic Regression

- The advantages of logistic regression are primarily the result of the general lack of assumptions.
- Logistic regression does not require any specific distributional form for the independent variables.
- Heteroscedasticity of the independent variables is not required.
- Linear relationships between the dependent and independent variables are not required.

Stage 4: Estimation of Logistic Regression Model and Assessing Overall Fit

- Transforming the dependent variable
- Estimating the coefficients
- Transforming a probability into odds and logit values
- Model estimation
- Assessing the goodness of fit

Estimating the Coefficients

Two basic steps . . .

1. Transforming a probability into odds and logit values

2. Model estimation using a maximum likelihood approach, not least squares as in multiple regression

   - The estimation process maximizes the likelihood that an event will occur – the event being a respondent is assigned to one group versus another

Transforming a Probability into Odds and Logit Values

- The logistic transformation has two basic steps:
  - ✓ Restating a probability as odds, and
  - ✓ Calculating the logit values.
- Instead of using ordinary least squares to estimate the model, the maximum likelihood method is used.
- The basic measure of how well the maximum likelihood estimation procedure fits is the likelihood value.

Model Estimation Fit – Between Model comparisons . . .

Comparisons of the likelihood values follow three steps:

1. Estimate a Null Model – which acts as the "baseline" for making comparisons of improvement in model fit.

2. Estimate Proposed Model – the model containing the independent variables to be included in the logistic regression.

3. Assess – 2LL Difference.

Comparison to Multiple Regression . . .

Correspondence of Primary Elements of Model Fit

Multiple Regression	Logistic Regression
Total Sum of Squares	-2LL of Base Model
Error Sum of Squares	-2LL of Proposed Model
Regression Sum of Squares Base	Difference of -LL for and Proposed Models
F test of model fit 2LL	Chi-square Test of - Difference
Coefficient of determination	“Pseudo” R^2 measures

Stage 5: Interpretation of the Results

- Testing for significance of the coefficients – based on the Wald statistic
- Interpreting the coefficients
- Directionality of the relationship
- Magnitude of the relationship of metric independent variables
- Interpreting nonmetric independent variables

Directionality of the Relationship

A positive relationship means an increase in the independent variable is associated with an increase in the predicted probability, and vice versa. But the direction of the relationship is reflected differently for the original and exponentiated logistic coefficients.

- Original coefficient signs indicate the direction of the relationship.

- Exponentiated coefficients are interpreted differently since they are the logarithms of the original coefficients and do not have negative values. Thus, exponentiated coefficients above 1.0 represent a positive relationship and values less than 1.0 represent negative relationships.

Magnitude of the Relationship . . .

The magnitude of metric independent variables is interpreted differently for original and exponentiated logistic coefficients:

- Original logistic coefficients – are less useful in determining the magnitude of the relationship since the reflect the change in the logit (logged odds) value.

- Exponentiated coefficients – directly reflect the magnitude of the change in the odds value. But their impact is multiplicative and a coefficient of 1.0 denotes no change (1.0 times the independent variable = no change).

Rules of Thumb 6–1

Logistic Regression

- Logistic regression is the preferred method for two-group (binary) dependent variables due to its robustness, ease of interpretation and diagnostics.
- Sample size considerations for logistic regression are primarily focused on the size of each group, which should have 10 times the number of estimated model coefficients (the number of variables).
- Sample size should be met in both the analysis and holdout samples.
- Model significance tests are made with a chi-square test on the differences in the log likelihood values (-2LL) between two models.

Rules of Thumb 6–1 continued . . .

Logistic Regression

- Coefficients are expressed in two forms: original and exponentiated to assist in interpretation.
- Interpretation of the coefficients for direction and magnitude is:
  - ✓ Direction can be directly assessed in the original coefficients (positive or negative signs) or indirectly in the exponentiated coefficients (less than 1 are negative, greater than 1 are positive).
  - ✓ Magnitude is best assessed by the exponentiated coefficient, with the percentage change in the dependent variable shown by: Percentage change = (Exponentiated Coefficient – 1.0) * 100

Stage 6: Validation of the Results

- Involves ensuring both the internal and external validity of the results.
- The most common form of estimating external validity is creation of a holdout or validation sample and calculating the hit ratio.
- A second approach is cross-validation, typically achieved with a jackknife or “leave-one-out” process of calculating the hit ratio.

TABLE 6-5 Calculating Estimated Probability Values for the Group Centroids of X_4 Region

	X_4 *(Region)*	
	Group 0: USA/North America	**Group 1: Outside North America**
Centroid: X_{13}	5.60	7.42
Centroid: X_{17}	3.63	4.93
Logit Value[a]	−1.452	2.909
Odds[b]	.234	18.332
Probability[c]	.189	.948

[a]Calculated as: $\text{Logit} = -14.190 + 1.079X_{13} + 1.844X_{17}$
[b]Calculated as: $\text{Odds} = e^{\text{Logit}}$
[c]Calculated as: $\text{Probability} = \text{Odds}/(1 + \text{Odds})$

Description of HBAT Primary Database Variables

	Variable Description	Variable Type
Data Warehouse Classification Variables		
X1	Customer Type	nonmetric
X2	Industry Type	nonmetric
X3	Firm Size	nonmetric
X4	Region	nonmetric
X5	Distribution System	nonmetric
Performance Perceptions Variables		
X6	Product Quality	metric
X7	E-Commerce Activities/Website	metric
X8	Technical Support	metric
X9	Complaint Resolution	metric
X10	Advertising	metric
X11	Product Line	metric
X12	Salesforce Image	metric
X13	Competitive Pricing	metric
X14	Warranty & Claims	metric
X15	New Products	metric
X16	Ordering & Billing	metric
X17	Price Flexibility	metric
X18	Delivery Speed	metric
Outcome/Relationship Measures		
X19	Satisfaction	metric
X20	Likelihood of Recommendation	metric
X21	Likelihood of Future Purchase	metric
X22	Current Purchase/Usage Level	metric
X23	Consider Strategic Alliance/Partnership in Future	nonmetric