What is Logit Regression and how is it used in Mplus data analysis?

Authored by

stats writer

June 29, 2024

Logit Regression is a statistical technique used to analyze categorical data and make predictions about the likelihood of an event occurring. In Mplus data analysis, it is used to model the relationship between a binary outcome variable and one or more predictor variables. This allows researchers to understand the influence of these predictors on the outcome variable and make inferences about the underlying population. Logit Regression is commonly used in Mplus to analyze survey data, evaluate the effectiveness of interventions, and predict the probability of a certain outcome based on various factors. It is a powerful tool for understanding the relationship between variables and making informed decisions based on data.

Logit Regression | Mplus Data Analysis Examples

Note: This example was done using Mplus version 6.12.

Logistic regression, also called a logit model, is used to model dichotomous outcome variables. In the logit model the log odds of the outcome is modeled as a linear combination of the predictor variables.

Please note: The purpose of this page is to show how to use various data analysis commands. It does not cover all aspects of the research process which researchers are expected to do. In particular, it does not cover data cleaning and checking, verification of assumptions, model diagnostics and potential follow-up analyses.

Examples

Example 1: Suppose that we are interested in the factors

that influence whether a political candidate wins an election. The

outcome (response) variable is binary (0/1); win or lose.

The predictor variables of interest are the amount of money spent on the campaign, the

amount of time spent campaigning negatively, and whether the candidate is an

incumbent.

Example 2: A researcher is interested in how variables, such as GRE (Graduate Record Exam scores),

GPA (grade

point average) and prestige of the undergraduate institution, effect admission into graduate

school. The outcome variable, admit/don't admit, is binary.

Description of the data

For our data analysis below, we are going to expand on Example 2 about getting into graduate school. We have generatedhypotheticaldata, which can be obtained by clicking on https://stats.idre.ucla.edu/wp-content/uploads/2016/02/binary.dat. You can store this anywhere you like, but our examples will assume it has been stored in c:data. (Note that the names of variables should NOT be included at the top of the data file. Instead, the variables are named as part of the variable command.) You may want to do your descriptive statistics in a general use statistics package, such as SAS, Stata or SPSS, because the options for obtaining descriptive statistics are limited in Mplus. Even if you chose to run descriptive statistics in another package, it is a good idea to run a model with type=basic before you do anything else,

just to make sure the dataset is being read correctly.

This dataset has data on 400 cases. There is a binary response (outcome, dependent) variable called admit and there are three predictor variables: gre, gpa, and rank. We will treat the variables gre and gpa as continuous. The variable rank takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest. The dataset also contains four dummy variables, one for each level of rank, named rank1 to rank4, for example, rank1 is equal to 1 when rank=1, and 0 otherwise.

Lets start by running a model with type=basic.

Data:

File is c:datahttps://stats.idre.ucla.edu/wp-content/uploads/2016/02/binary.dat ;

Variable:

Names are
admit gre gpa rank rank1 rank2 rank3 rank4;
Analysis:
Type = basic ;

As we mentioned above, you will want to look at this carefully to be sure that
the dataset was read into Mplus correctly. You will want to make sure that
you have the correct number of observations, and that the variables all have
means that are close to those from the descriptive statistics generated in a
general purpose statistical package. If there are missing values for some or all
of the variables, the descriptive statistics generated by Mplus will not match
those from a general purpose statistical package exactly, because by default, Mplus versions
5.0 and later use maximum likelihood based procedures for handling missing
values. The main point of running this model is to make sure that the data is
being read correct by Mplus, if the number of cases and

variables is correct,
and the means are reasonable, then it is probably safe to proceed.

<output omitted>

SUMMARY OF ANALYSIS

Number of groups 1
Number of observations 400

<output omitted>

SAMPLE STATISTICS

Means
ADMIT GRE GPA RANK RANK1

_____ _____ _____ _____ _____

1 0.318 587.700 3.390 2.485 0.152

Means
RANK2 RANK3 RANK4

_____ _____ _____

1 0.378 0.302 0.168

Analysis methods you might consider

Below is a list of some analysis methods you may have encountered.
Some of the methods listed are quite reasonable while others have either
fallen out of favor or have limitations.

Using the logit model

The Mplus input file for a logistic regression model is shown below. Because
the data file contains variables that are not used in the model, the usevariables subcommand
is used to list the variables that appear in the model (i.e., admit,
gre, gpa, rank1, rank2, and rank3). Note that because Mplus uses the names subcommand to determine the order of variables
in the data file, the number and order of variables in the names subcommand should not be changed unless the data file is also changed. The
categorical subcommand is used to identify binary and ordinal outcome
variables. Only the categorical outcome

variable (i.e., admit) is included in the categorical subcommand. Categorical predictor variables should be included as a series of dummy variables (e.g., rank1, rank2, and rank3). Under analysis we have specified estimator=ml, this requests a logit model, rather than the default probit model. Finally, in the model command we specify that the outcome (i.e., admit) should be regressed on the predictor variables (i.e., gre, gpa, rank1, rank2, and rank3).

Data:
File is c:datahttps://stats.idre.ucla.edu/wp-content/uploads/2016/02/binary.dat ;
Variable:
names = admit gre gpa rank rank1 rank2 rank3 rank4;
usevariables = admit gre gpa rank1 rank2 rank3;
categorical = admit;
Analysis:
estimator = ml;

Model:

admit on gre gpa rank1 rank2 rank3;

SUMMARY OF ANALYSIS

Number of groups 1

Number of observations 400

Number of dependent variables 1

Number of independent variables 5

Number of continuous latent variables 0

Observed dependent variables

Binary and ordered categorical (ordinal)

ADMIT

Observed independent variables

GRE GPA RANK1 RANK2 RANK3

Estimator ML

Information matrix OBSERVED

Optimization Specifications for the Quasi-Newton Algorithm for

Continuous Outcomes

Maximum number of iterations 100

Convergence criterion 0.100D-05

Optimization Specifications for the EM Algorithm

Maximum number of iterations 500

Convergence criteria

Loglikelihood change 0.100D-02

Relative loglikelihood change 0.100D-05

Derivative 0.100D-02

Optimization Specifications for the M step of the EM Algorithm for

Categorical Latent variables

Number of M step iterations 1

M step convergence criterion 0.100D-02

Basis for M step termination ITERATION

Optimization Specifications for the M step of the EM Algorithm for

Censored, Binary or Ordered Categorical (Ordinal), Unordered

Categorical (Nominal) and Count Outcomes

Number of M step iterations 1

M step convergence criterion 0.100D-02

Basis for M step termination ITERATION

Maximum value for logit thresholds 15

Minimum value for logit thresholds -15

Minimum expected cell size for chi-square 0.100D-01

Optimization algorithm EMA

Integration Specifications

Type STANDARD

Number of integration points 15

Dimensions of numerical integration 0

Adaptive quadrature ON

Link LOGIT

Cholesky OFF

Input data file(s)

C:datahttps://stats.idre.ucla.edu/wp-content/uploads/2016/02/binary.dat

Input data format FREE

SUMMARY OF CATEGORICAL DATA PROPORTIONS

ADMIT

Category 1 0.683

Category 2 0.317

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Loglikelihood

H0 Value -229.259

Information Criteria

Number of Free Parameters 6
Akaike (AIC) 470.517
Bayesian (BIC) 494.466
Sample-Size Adjusted BIC 475.428
(n* = (n + 2) / 24)

MODEL RESULTS

Two-Tailed
Estimate S.E. Est./S.E. P-Value

ADMIT ON
GRE 0.002 0.001 2.070 0.038
GPA 0.804 0.332 2.423 0.015
RANK1 1.551 0.418 3.713 0.000
RANK2 0.876 0.367 2.389 0.017
RANK3 0.211 0.393 0.538 0.591

Thresholds
ADMIT$1 5.541 1.138 4.869 0.000

LOGISTIC REGRESSION ODDS RATIO RESULTS

ADMIT ON
GRE 1.002
GPA 2.235
RANK1 4.718
RANK2 2.401
RANK3 1.235

We can also test that the coefficients for rank1, rank2, and rank3, are all equal to zero using the model test command. This type of test can also be described as an overall test for the effect of rank. There are multiple ways to test this type of hypothesis, the model test command requests a Wald test. The Mplus input file shown below is similar to the first model, except that the coefficients for rank1, rank2, and rank3 are assigned the names r1, r2, and r3, respectively. In the model test command, these coefficient names (i.e., r1, r2 and r3) are used to

test that each of the coefficients is equal to 0.

Data:
File is C:datahttps://stats.idre.ucla.edu/wp-content/uploads/2016/02/binary.dat ;
Variable:
names = admit gre gpa rank rank1 rank2 rank3 rank4;
categorical = admit;
usevariables = admit gre gpa rank1 rank2 rank3;
Analysis:
estimator = ML;
Model:
admit on gre gpa
rank1 (r1)
rank2 (r2)
rank3 (r3);
Model test:
r1 = 0;
r2 = 0;
r3 = 0;

The majority of the output from this model is the same as the first model, so we will only show part of

the output generated by the model test command.

TESTS OF MODEL FIT

Wald Test of Parameter Constraints

Value 20.895
Degrees of Freedom 3
P-Value 0.0001

Loglikelihood

H0 Value -229.259

The portion of the output associated with the model test command is labeled "Wald Test of Parameter Constraints" and appears under the heading TESTS OF MODEL FIT just before the likelihood for the entire model is printed. The test statistic is 20.895, with three degrees of freedom (one for each of the parameters tested), with an associated p-value of 0.0001. This indicates that the overall effect of rank is statistically significant.

We can also use the model test command to make pairwise comparisons among the terms for rank. The Mplus input below tests the hypothesis that the coefficient for rank2 (i.e., rank=2) is equal to the coefficient for rank3 (i.e., rank=3).

Data:
File is C:datahttps://stats.idre.ucla.edu/wp-content/uploads/2016/02/binary.dat ;
Variable:
names = admit gre gpa rank rank1 rank2 rank3 rank4;
categorical = admit;
usevariables = admit gre gpa rank1 rank2 rank3;
Analysis:
estimator = ml;
Model:
admit on gre gpa
rank1 (r1)
rank2 (r2)
rank3 (r3);
Model test:

r2 = r3;

**Below is the output associated with the model test command (as before,
most of the model output is omitted).**

MODEL FIT INFORMATION

Wald Test of Parameter Constraints

**Value 5.505
Degrees of Freedom 1
P-Value 0.0190**

**The test statistic and associated p-value indicate that the coefficient for rank2 (i.e., rank=2)
is significantly different from the coefficient for rank3 (rank=3).**

Things to consider

References

**Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition).
New York: John Wiley & Sons, Inc.**

Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

See also