# What is Multinomial Logistic Regression and how can it be used in SAS data analysis?

Authored by

**stats writer**

June 29, 2024

Multinomial Logistic Regression is a statistical technique used to model and predict the relationship between multiple categorical dependent variables and one or more independent variables. It is an extension of binary logistic regression, which is commonly used for binary outcomes.

In SAS data analysis, Multinomial Logistic Regression can be used to analyze data with multiple outcome categories. This can be helpful in situations where there are more than two possible outcomes, such as predicting the likelihood of a customer purchasing one of several products or predicting the success of different treatment options in a medical study.

The model estimates the probability of each outcome category based on the independent variables, allowing for the identification of significant predictors and the comparison of the effects of different variables on each outcome. This can provide valuable insights for decision making and can be useful in various fields such as marketing, healthcare, and social sciences.

Furthermore, SAS offers various tools and procedures for Multinomial Logistic Regression analysis, making it easily accessible and customizable for different data sets and research purposes. It is a powerful and widely used tool in data analysis, providing a flexible and efficient way to examine relationships between categorical variables.



## Multinomial Logistic Regression | SAS Data Analysis Examples

## Version info: Code for this page was tested in SAS 9.3.

## Multinomial logistic regression is for modeling nominal outcome variables, in which the log odds of the outcomes are 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 of multinomial logistic regression

Example 1. People's occupational choices might be influenced
by their parents' occupations and their own education level. We can study the
relationship of one's occupation choice with education level and father's
occupation. The occupational choices will be the outcome variable which
consists of categories of occupations.

Example 2. A biologist may be interested in food choices that alligators make. Adult alligators might have
difference preference than young ones. The outcome variable here will be the
types of food, and the predictor variables might be the

length of the alligators
and other environmental variables.

Example 3. Entering high school students make program choices among general program,
vocational program and academic program. Their choice might be modeled using
their writing score and their social economic status.

### Description of the data

For our data analysis example, we will expand the third example using the
hsbdemo data set. You can download the data
here .

```
proc contents data = "c:hsbdemo";
run;
```

The CONTENTS Procedure

Data Set Name c:datahsbdemo Observations 200
Member Type DATA Variables 13
Engine V9 Indexes 0
Created Thursday, August 29, 2013 09:42:59 AM
Observation Length 40

**Last Modified Thursday, August 29, 2013 09:42:59 AM**
**Deleted Observations 0**
**Protection Compressed NO**
**Data Set Type Sorted YES**
**Label Written by SAS**

**Data Representation WINDOWS_64**
**Encoding wlatin1 Western (Windows)**

**Engine/Host Dependent Information**

**Data Set Page Size 4096**
**Number of Data Set Pages 3**
**First Data Page 1**
**Max Obs per Page 101**
**Obs in First Data Page 42**
**Number of Data Set Repairs 0**
**Filename c:datahsbdemo.sas7bdat**
**Release Created 9.0301M1**
**Host Created X64_7PRO**

**Alphabetic List of Variables and Attributes**

**# Variable Type Len Label**

**12 AWARDS Num 3**

13 CID Num 3

2 FEMALE Num 3

11 HONORS Num 3 honores eng

1 ID Num 4

8 MATH Num 3 math score

5 PROG Num 3 type of program

6 READ Num 3 reading score

4 SCHTYP Num 3 type of school

9 SCIENCE Num 3 science score

3 SES Num 3

10 SOCST Num 3 social studies score

7 WRITE Num 3 writing score

Sort Information

Sortedby PROG

Validated YES

Character Set ANSI

The data set contains variables on 200 students. The outcome variable is prog, program type. The predictor variables
are social economic status, ses, a three-level categorical variable

**and writing score, write, a continuous variable. Let's start with**
**getting some descriptive statistics of the**
**variables of interest.**

```
proc freq data = "c:hsbdemo";
tables prog*ses / chisq norow nocol nofreq;
run;
```

**The FREQ Procedure**

**Table of PROG by SES**

**PROG(type of program) SES**

| Percent | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| 1 | 8.00 | 10.00 | 4.50 | 22.50 |
| 2 | 9.50 | 22.00 | 21.00 | 52.50 |
| 3 | 6.00 | 15.50 | 3.50 | 25.00 |
| Total | 47 | 95 | 58 | 200 |
| | 23.50 | 47.50 | 29.00 | 100.00 |

## Statistics for Table of PROG by SES

## Statistic DF Value Prob

---

## Chi-Square 4 16.6044 0.0023

## Likelihood Ratio Chi-Square 4 16.7830 0.0021

## Mantel-Haenszel Chi-Square 1 0.0598 0.8068

## Phi Coefficient 0.2881

## Contingency Coefficient 0.2769

## Cramer's V 0.2037

## Sample Size = 200

```
proc sort data = "c:hsbdemo";
by prog;
run;

proc means data = "c:hsbdemo";
var write;
by prog;
run;
```

## type of program=1

## The MEANS Procedure

## Analysis Variable : WRITE writing score

## N Mean Std Dev Minimum Maximum

---

## 45 51.3333333 9.3977754 31.0000000 67.0000000

---

## type of program=2

## Analysis Variable : WRITE writing score

## N Mean Std Dev Minimum Maximum

---

## 105 56.2571429 7.9433433 33.0000000 67.0000000

---

## type of program=3

## Analysis Variable : WRITE writing score

## N Mean Std Dev Minimum Maximum

---

## 50 46.7600000 9.3187544 31.0000000 67.0000000

---

## Analysis methods you might consider

## Multinomial logistic regression

**Below we use proc logistic to estimate a multinomial logistic
regression model. The outcome prog and the predictor ses are both
categorical variables and should be indicated as such on the class statement. We
can specify the baseline category for prog using (ref = "2") and
the reference group for ses using (ref = "1"). The param=ref option
on
the class statement tells SAS to use dummy coding rather than effect coding
for the variable ses. Note that the levels of prog are defined as:**

**1=general**

**2=academic (reference group)**

**3=vocational**

**proc logistic data = "c:hsbdemo";**
**class prog (ref = "2") ses (ref = "1") / param = ref;**
**model prog = ses write / link = glogit;**
**run;**
**The LOGISTIC Procedure**

**Model Information**

**Data Set c:datahsbdemo Written by SAS**
**Response Variable PROG type of program**
**Number of Response Levels 3**
**Model generalized logit**
**Optimization Technique Newton-Raphson**

**Number of Observations Read 200**
**Number of Observations Used 200**

**Response Profile**

**Ordered Total**
**Value PROG Frequency**

**1 1 45**
**2 2 105**
**3 3 50**

## Logits modeled use PROG=2 as the reference category.

## Class Level Information

## Design

## Class Value Variables

## SES 1 0 0

## 2 1 0

## 3 0 1

## Model Convergence Status

## Convergence criterion (GCONV=1E-8) satisfied.

## Model Fit Statistics

## Intercept

## Intercept and

## Criterion Only Covariates

## AIC 412.193 375.963

## SC 418.790 402.350

## -2 Log L 408.193 359.963

## Testing Global Null Hypothesis: BETA=0

**Test Chi-Square DF Pr > ChiSq**

**Likelihood Ratio 48.2299 6 <.0001**
**Score 45.1588 6 <.0001**
**Wald 37.2946 6 <.0001**

**Type 3 Analysis of Effects**

**Wald**
**Effect DF Chi-Square Pr > ChiSq**

**SES 4 10.8162 0.0287**
**WRITE 2 26.4633 <.0001**

**Analysis of Maximum Likelihood Estimates**

**Standard Wald**
**Parameter PROG DF Estimate Error Chi-Square Pr > ChiSq**

**Intercept 1 1 2.8522 1.1664 5.9790 0.0145**
**Intercept 3 1 5.2182 1.1635 20.1128 <.0001**
**SES 2 1 1 -0.5333 0.4437 1.4444 0.2294**
**SES 2 3 1 0.2914 0.4764 0.3742 0.5407**
**SES 3 1 1 -1.1628 0.5142 5.1137 0.0237**
**SES 3 3 1 -0.9827 0.5956 2.7224 0.0989**

**WRITE 1 1 -0.0579 0.0214 7.3200 0.0068**
**WRITE 3 1 -0.1136 0.0222 26.1392 <.0001**

**Odds Ratio Estimates**

**Point 95% Wald**
**Effect PROG Estimate Confidence Limits**

**SES 2 vs 1 1 0.587 0.246 1.400**
**SES 2 vs 1 3 1.338 0.526 3.404**
**SES 3 vs 1 1 0.313 0.114 0.856**
**SES 3 vs 1 3 0.374 0.116 1.203**
**WRITE 1 0.944 0.905 0.984**
**WRITE 3 0.893 0.855 0.932**

**Two models are tested in this multinomial regression, one comparing**
**membership to general versus academic program and one comparing membership to**
**vocational versus academic program. They correspond to the two equations below:**

**$$ln\left(\frac{P(prog=general)}{P(prog=academic)}\right) = b_{10} + b_{11}(ses=2) + b_{12}(ses=3) + b_{13}write$$**

$$ln\left(\frac{P(prog=vocation)}{P(prog=academic)}\right) = b_{20} + b_{21}(ses=2) + b_{22}(ses=3) + b_{23}write$$

where (b)s are the regression coefficients.

Using the test statement, we can also test specific hypotheses within
or even across logits, such as if the effect of ses=3 in predicting general versus academic equals the effect of ses = 3 in
predicting vocational versus academic. Use of the test statement requires the
unique names SAS assigns each parameter in the model. The option outest
on the proc logistic statement produces an output dataset with
the parameter names and values. We can get these names by printing them,
and we transpose them to be more readable. The noobs option on the proc print
statement suppresses observation numbers, since they are meaningless in the parameter dataset.

```
proc logistic data = "c:hsbdemo" outest = mlogit_param;
class prog (ref = "2") ses (ref = "1") / param = ref;
model prog = ses write / link = glogit;
run;proc transpose data = mlogit_param;
run;
proc print noobs;
run;
```

_NAME_ _LABEL_ PROG

Intercept_1 Intercept: PROG=1 2.852
Intercept_3 Intercept: PROG=3 5.218
SES2_1 SES 2: PROG=1 -0.533
SES2_3 SES 2: PROG=3 0.291
SES3_1 SES 3: PROG=1 -1.163
SES3_3 SES 3: PROG=3 -0.983
WRITE_1 writing score: PROG=1 -0.058
WRITE_3 writing score: PROG=3 -0.114
_LNLIKE_ Model Log Likelihood -179.982

Here we see the same parameters as in the output above, but with their unique SAS-given names.
We are interested in testing whether SES3_general is equal to SES3_vocational,

**which we can now do with the test statement. The code preceding the ":"**
**on the test statement is a label identifying the test in the output, and it must**
**conform to SAS variable-naming rules (i.e., 32 characters in length or less, letters,**
**numerals, and underscore).**

```
proc logistic data = "c:hsbdemo" outest = mlogit_param;
class prog (ref = "2") ses (ref = "1") / param = ref;
model prog = ses write / link = glogit;
SES3_general_vs_SES3_vocational: test SES3_1 - SES3_3;
run;
```

*****SOME OUTPUT OMITTED*****

**Linear Hypotheses Testing Results**

**Wald**
**Label Chi-Square DF Pr > ChiSq**

**SES3_general_vs_SES3_vocational 0.0772 1 0.7811**

**The effect of ses=3 for predicting general versus academic is not different from the effect of ses=3 for predicting vocational versus academic.**

**You can also use predicted probabilities to help you understand the model.**
**You can calculate predicted probabilities using the lsmeans statement and the ilink option. For multinomial data, lsmeans requires glm rather than reference (dummy) coding, even though they are essentially the same, so be sure to respecify the coding on the class statement.**
**However, glm coding only allows the last category to be the reference group (prog = vocational and ses = 3)and will ignore any other reference group specifications. Below we use lsmeans to calculate the predicted probability of choosing program type academic or general at each level of ses, holding write at its means.**

```
proc logistic data = "c:hsbdemo" outest =
mlogit_param;
class prog ses / param = glm;
model prog = ses write / link = glogit;
lsmeans ses / e ilink cl;
run;
```

***SOME OUTPUT OMITTED***

Coefficients for SES Least Squares Means

type of
Parameter program SES Row1 Row2 Row3 Row4 Row5 Row6

Intercept 1 1 1 1
Intercept 2 1 1 1
SES 1 1 1 1
SES 1 2 1 1
SES 2 1 2 1
SES 2 2 2 1
SES 3 1 3 1
SES 3 2 3 1
writing score 1 52.775 52.775 52.775
writing score 2 52.775 52.775 52.775

*****SOME OUTPUT OMITTED*****

**SES Least Squares Means**

| type of program | SES | Standard Error of Mean | Lower Mean | Upper Mean |
|---|---|---|---|---|
| 1 | 1 | 0.3582 | 0.07264 | 0.2158 | 0.5006 |
| 1 | 2 | 0.2283 | 0.04512 | 0.1399 | 0.3168 |
| 1 | 3 | 0.1785 | 0.05405 | 0.07256 | 0.2844 |
| 2 | 1 | 0.4397 | 0.07799 | 0.2868 | 0.5925 |
| 2 | 2 | 0.4777 | 0.05526 | 0.3694 | 0.5861 |
| 2 | 3 | 0.7009 | 0.06630 | 0.5709 | 0.8309 |

**The predicted probabilities are in the "Mean" column. Thus, for ses = 3 and write = 52.775, we see that the probability of being the academic program (program type 2) is 0.7009; for the general program (program type 1), the probability is 0.1785.**

**To obtain predicted probabilities for the program type vocational, we can reverse the ordering of the**

**categories**
**using the descending option on the proc logistic statement.**
**This will make academic the reference group for prog and 3 the reference**
**group for ses.**

```
proc logistic data = "c:hsbdemo" outest = mlogit_param descending;
class prog ses / param = glm;
model prog = ses write / link = glogit;
lsmeans ses / e ilink cl;
run;
```

*****SOME OUTPUT OMITTED*****
**Coefficients for SES Least Squares Means**

**type of**
**Parameter program SES Row1 Row2 Row3 Row4 Row5 Row6**

**Intercept 3 1 1 1**
**Intercept 2 1 1 1**
**SES 1 3 1 1**
**SES 1 2 1 1**

**SES 2 3 2 1**

**SES 2 2 2 1**

**SES 3 3 3 1**

**SES 3 2 3 1**

**writing score 3 52.775 52.775 52.775**

**writing score 2 52.775 52.775 52.775**

*****SOME OUTPUT OMITTED*****

**SES Least Squares Means**

**Standard**

**type of Error of Lower Upper**

**program SES Mean Mean Mean Mean**

**3 1 0.2021 0.05996 0.08459 0.3197**

**3 2 0.2939 0.05036 0.1952 0.3926**

**3 3 0.1206 0.04643 0.02960 0.2116**

**2 1 0.4397 0.07799 0.2868 0.5925**

**2 2 0.4777 0.05526 0.3694 0.5861**

**2 3 0.7009 0.06630 0.5709 0.8309**

**Here we see the probability of being in the vocational program when ses = 3 and**

**write = 52.775 is 0.1206, which is what we would have**

**expected since (1 -
0.7009 - 0.1785) = 0.1206, where 0.7009 and 0.1785 are the probabilities of
being in the academic and general programs under the same conditions.**

## Things to consider

## See Also

## References