

# Logistic regression analysis

2019.04.10

## Introduction

In the linear regression model, the dependent variable is a quantitative variable, not a qualitative variable. However, in many practical problems, it is often the case that the dependent variable is a qualitative variable. Logistic regression analysis, Probit analysis, log-linear model, etc. are statistical methods for processing classification dependent variables. The Logistic regression model is a multivariate analysis method for analyzing and predicting discrete dependent variables based on single or multiple continuous or discrete independent variables. It is a commonly used statistical method.

This article will present an example of a logistic regression model for analysis and a brief analysis and evaluation.

## Logistic

Logistic regression analysis, Probit analysis, discriminant analysis, log-linear model, etc. are statistical methods for processing classification dependent variables. Among them, Logistic regression analysis, Probit analysis is a commonly used statistical method for processing classification dependent variables. The Logistic regression model is a multivariate analysis method for analyzing and predicting discrete dependent variables based on single or multiple continuous or discrete independent variables. Logistic regression analysis is the most commonly used statistical method.

The Logistic probability function, also known as the growth function, was first proposed by P.F.V erhuis in Belgium in 1838. After using the logistic function as the growth curve, he used the curve for demographic research until the end of the 19th century. Logistic regression models have been used extensively for many years. At the time of its origin, it has been applied in the study of infectious diseases. Logistic regression analysis has been widely used in many fields as an effective data processing method, such as biomedicine, criminology, ecological engineering, health, linguistics and wildlife, biology.

## Application

This paper uses logistic model regression to analyze a practical problem. The problem description is: the milk products currently sold in supermarkets mainly

include Mengniu, Yili, Guangming and Manor. Now it is necessary to analyze the different ages and wages, and whether it will affect the choice of milk products brand.

The data of this regression analysis is randomly generated by EXCEL, where the “choice” variable is the interpreted variable, indicating the brand of the milk selected by the customer; the explanatory variables include “age” and “income”, indicating the age and monthly income of the customer, respectively. For the “choice” variable, the values “1”, “2”, “3”, and “4” represent the brands of Mengniu, Yili, Guangming, and Manor, respectively. (Specific data can be seen in the notes.)

We perform Logit regression analysis in STATA, which can handle multi-valued individual-specific explanatory variables. The code and results are as follows:

```
mlogit choice age income
predict c1 c2 c3 c4 in 1/10
. mlogit choice age income

Iteration 0:   log likelihood = -132.58649
Iteration 1:   log likelihood = -129.91023
Iteration 2:   log likelihood = -129.90474
Iteration 3:   log likelihood = -129.90474

Multinomial logistic regression               Number of obs   =           100
                                                LR chi2(6)       =           5.36
                                                Prob > chi2      =          0.4981
Log likelihood = -129.90474                   Pseudo R2       =          0.0202
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
1	(base outcome)					
2						
age	.1159301	.1317833	0.88	0.379	-.1423605	.3742206
income	-.005675	.0050338	-1.13	0.260	-.0155409	.004191
_cons	18.00942	20.70307	0.87	0.384	-22.56785	58.5867
3						
age	.2314693	.1408797	1.64	0.100	-.0446499	.5075885
income	.0005456	.0055545	0.10	0.922	-.010341	.0114321
_cons	-11.76394	22.98534	-0.51	0.609	-56.81438	33.2865
4						
age	.002081	.1788722	0.01	0.991	-.3485021	.352664
income	.0019682	.0069224	0.28	0.776	-.0115994	.0155359
_cons	-8.980221	28.63694	-0.31	0.754	-65.10759	47.14715

The predicted results are shown in the figure below. (Detailed data can be seen in the note.)

	age	income	id	choice	c1	c2	c3	c4
1	43	4031	1	2	.2509485	.2825712	.3700794	.0964009
2	39	4007	2	1	.3634287	.2949319	.2095751	.1320644
3	41	3919	3	2	.2586531	.4361147	.2258598	.0793724
4	42	4072	4	3	.2935456	.2332505	.3512164	.1219875
5	40	4130	5	3	.3711109	.1682706	.2884656	.1721529
6	37	4064	6	2	.4454358	.2074483	.166786	.1803299
7	38	4047	7	4	.4105323	.2364385	.1919639	.1610653
8	40	4025	8	1	.3408215	.2804236	.2501718	.128583
9	41	4064	9	1	.3232263	.239344	.3054812	.1319484
10	36	3926	10	2	.4014423	.3643469	.1106047	.123606

Of course, we can also use the "asclogit" command for regression analysis. Compared to the "mlogit" command, it can handle not only the individual-specific explanatory variables, but also the alternative-specific explanatory variables.

In addition, "mlogit" and "mprobit" handle wide data, while "clogit" handles long data. It is not shown here. Readers can try to use these regression analysis to solve problems in stata.

## References

- [1] Yin Jianjie. Summary and Applied Research on Regression Model Analysis[D]. Heilongjiang:Heilongjiang University,2011.
- [2] Gumbel, E. J. Ranges and Midranges[J]. Ann. Math. Statist, 1994, (15): 300-322.