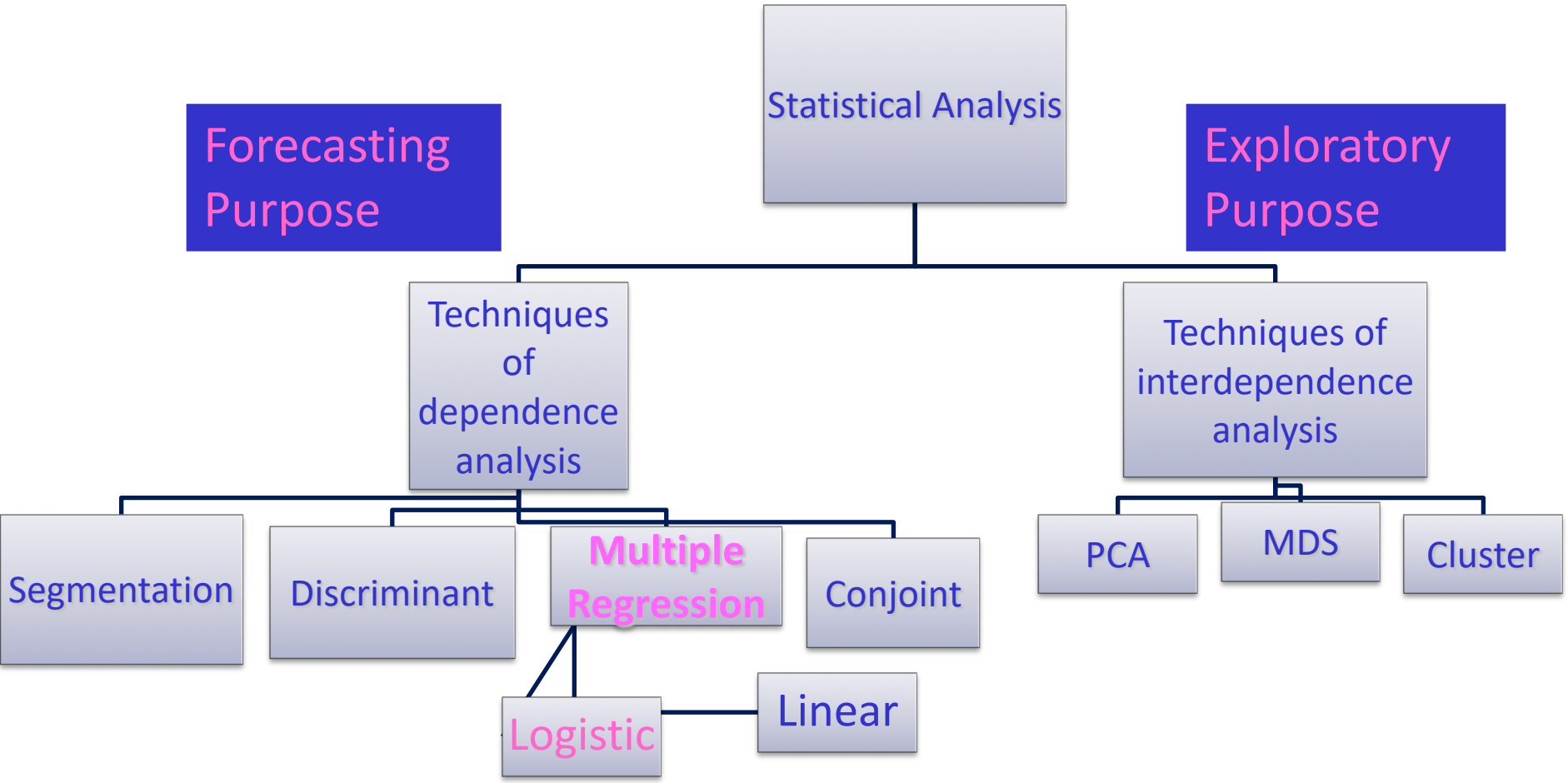
Main quantitative analysis techniques



Logistic Regression Model

Aim of the logistic regression

The basic idea of logistic regression (defined as *logit*) is the creation of a non linear model which identifies the main characteristics with the aim to forecast sales, to measure the potential market, to study consumer behavior, consumer satisfaction and to perform a market segmentation.

Aim of the logistic regression

The innovation is the use of a qualitative dependent variable, dichotomous, taking value 1 if successful, 0 failure.

eg.: "buy" – "do not buy"

Encoded, respectively, with 1 (= "buy") and 0 (= " do not buy")

Logistic regression model allows to estimate the probability that the dependent variable can take one of two extreme values (usually 1), instead of the value of the variable

Example:

Estimate the probability (P) to occur the purchase of a product, rather than adopt a linear regression model to assess and predict product sales.

Technical problems

Dependent variables which assume only two values: 0 e 1;

while the corresponding regression function can assume all values in $[-\infty, +\infty]$

Logistic regression model

The *logit* model is defined by (in a multiple case):

$$P(Y=1 \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}} + \varepsilon$$

where Y is a bernoulli random variable with P(Y=1|X)

Given the explanatory variables, Xj, and the logistic random variable cdf (exp(Xβ)/1+exp(Xβ)), we can rewrite the previous equation as

$$P(Y=1) = \frac{e^{\sum_{j=0}^{k} \beta_j x_j}}{1 + e^{\sum_{j=0}^{k} \beta_j x_j}} + \varepsilon \qquad x_0 = 1$$

Logistic regression model

Explanatory variables can be both quantitative, assuming all values of real number interval, and *dummy variables*, which are encoded in numerical values of 0 and 1 (dichotomous).

Transformation of P in *odds*

In order to estimate the probability that Y takes a given value, and since

$$0 \leq \quad P \leq \quad 1$$

It can be useful to transform the probabilities in likelihood ratios, to obtain values greater than 1

odds

The "*odds*" is the ratio among the probability of a event P and its complement to 1:

$$odds = \frac{P}{1-P}$$

Transformation of *odds* in *logit*

A logarithmic transformation of the *odds, defined "logit"*, allows to get dependent variable values greater than one and lower than zero (negative).

$$odds = \frac{P}{1-P}$$

$$Logit(P) = \log_e(odds) = \log_e(P/1-P)$$

Logit model

Remembering that:

$$1 - P(Y = 1 | X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}} + \varepsilon$$

The *odds* is:

$$odds = \frac{P(Y=1)}{P(Y=0)} = \frac{P}{1-P} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k} = e^{\sum_{j=0}^{k} \beta_j x_j} = e^{x\beta} + \varepsilon$$

Then, the LOGIT (P) allows to get a linear regression model:

$$Logit(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon'$$

Parameter interpretation

$$Logit(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon'$$

β_j can be interpreted as the *logit* variation due to the unite increase of the explanatory variable x_j

- If does not exist relationship between x_j and the probability that Y is 1, $\beta_j=0$
- If there exist a positive relationship between x_j and the probability that Y is 1, $\beta_j>0$
- If there exist a negative relationship between x_j and the probability that Y is 1, $\beta_j<0$

Parameter estimation method

The main hypothesis is the absence of multicollinearity .

The method used to estimate the parameter vector is the maximum likelihood method.

This method is based on the maximization of the known likelihood function that maximizes the probability of observing the set of sample data in function of β.

Since:

$$y \sim bernulliana(p)$$

the probability function for the i-th observation is:

$$f(y_i \mid x_i; \beta) = P^{y_i}(1-P)^{1-y_i}$$

Parameter estimation

The log-likelihood of the observed sample (of size n), "L(β)", is given by the product of all log-likelihoods corresponding to sample units and it is function of β parameters

$$L(\beta) = \prod_{i=1}^{n} f(y_i \mid x_i; \beta) = \prod_{i=1}^{n} P^{y_i} (1-P)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{e^{x'_i \beta}}{1+e^{x_i \beta}} \right)^{y_i} \left(1 - \frac{e^{x_i \beta}}{1+e^{x_i \beta}} \right)^{1-y_i}$$

To obtain the maximum likelihood parameter vector β, we estimate the β values which maximize the logarithm of L(β).

To maximize the expression, we impose the partial derivative, with respect to the parameters, equal to zero and, then, an iterative estimation procedure is required.

Goodness of fit

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no relationship)
H_1: $\beta_1 \neq \beta_2 \neq \ldots \neq \beta_k \neq 0$ (at least one coefficient is different from zero)

$$G = dev(\text{modello}\,\beta_0) - dev(\text{modello completo}) = -2\log\frac{L(0)}{L(\beta)} = -2\log LR$$

This statistical test is the difference between the deviance of the model with only the intercept and the deviance of the model analysed.

LR (Likelihood Ratio) assumes values near 1 when the explanatory variables of the model are not significant and G is near 0; on the contrary, LR assumes values near 0 if the variables are significant, G value is high.

Goodness of fit

It has been showed that:

$$G \sim \chi^2_{k-1}$$

Where k= number of explanatory variables

If G is greater of critical value, H0 is rejected (with a given α level)

Goodness of fit

Cox and Snell index

$$R^2 = 1 - \left[\frac{L(0)}{L(\beta)}\right]^{2/n}$$

It ranges between 0 (if the estimated model does not provide information than the intercept-only one) and its maximum $R_{\max}{}^2 = 1 - [L(0)]^{2/n}$

To make the index moving between 0 and 1:

Nagelkerke's index

$$\overline{R}^2 = R^2 / R_{\max}{}^2$$

Wald's test

H_0: $\beta_j = 0$
H_1: $\beta_j \neq 0$

We use *Wald's test (W)*:

$$W = \left[\frac{b_j}{S_{b_j}} \right]^2$$

→ Estimation of β

→ Standard error of b

In multivariate case:

$$W = B'V^{-1}B$$

→ Maximum likelihood estimation vector of β

→ Inverse of covariance matrix of coefficients

$$W \sim \chi_1^2$$

If greater than critical value of χ^2 (with given α) H_0 is rejected

Example: probability of purchasing

To estimate the probability of buying a snack (1=yes, 0=no):

- Average number of pieces purchased in a month
- Age of respondents
- Exposure to advertising (*dummy*: 1= yes, I have seen the tv advertisement, 0=no)

We can use the following logistic regression model

$$Logit(P(purchase = 1) = \beta_0 + \beta_1(average\ pieces) + \beta_2(Age) + \beta_3(advertisement) + \varepsilon$$

Model with constant

Iteration History[a,b,c]

Iteration	-2 Log likelihood	Coefficients: Constant
Step 0 1	41.18	,000

a. Constant is included in the model.

$-2\log L(0)$

Goodness of fit

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	25.78	0.3821	0.5278

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

$$-2\log L(\beta)$$

$$R^2 = 1 - \left[\frac{L(0)}{L(\beta)}\right]^{2/n}$$

$$\overline{R}^2 = \frac{R^2}{R^2_{\max}}$$

Values high enough that confirm the significance of the model adopted!

Significance of all parameters

$$G = -2\log\frac{L(0)}{L(\beta)} = -2\log L(0) - (-2\log L(\beta))$$

$$G = 41.18 - 25.78 = 15.40$$

$$\chi^2 = 7.81$$

$$G > \chi^2 \text{ rejection of } H_0$$

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	15.40	3	0.0015
	Block	15.40	3	0.0015
	Model	15.40	3	0.0015

Sig values lead to rejection of H0,

K parameters = 4

K-1 explanatory variables = 3

Significance of single parameter

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Average	2.826	1.263	5.007	1	0.025	16.878
	Age	0.095	0.142	0.452	1	0.50	1.100
	Advertising	2.379	1.065	4.993	1	0.025	10.794
	Constant	-13.020	4.931	6.972	1	0.008	0.000

a. Variable(s) entered on step 1: pezziacquistati, eta, effettopubbl.

$$Logit(P(purchase = 1)) = -13{,}020 + 2{,}826(average\ pieces) + 0{,}095(Age) + 2{,}379(advertising)$$

Wald's test states the significance of all parameters, except age (Sig >0.05).

Exp(β) is the increase of *odds –ratio* with respect to unit increases of Xj

$EXP(B) = e^b$

e.g.:

$$odds = e^{-13,020+2,826(2)+0,095(22)+2,379(1)}$$

For unit increments of age, from 22 to 23 years

$$odds = e^{-13,020+2,826(2)+0,095(23)+2,379(1)}$$

The odds-ratio is

$$odds-ratio = \frac{odds(23)}{odds(22)} = \frac{e^{-13,020+2,826(2)+0,095(23)+2,379(1)}}{e^{-13,020+2,826(2)+0,095(22)+2,379(1)}} = e^{0,095(23-22)} = e^{0,095}$$

e^b

The b value states that a unit change of age has a positive impact on both, the odds-ratio (>1) and the probability of y=1.

RULE:

If X increases of a single unit, the odds-ratio may be equal to:

1 = does not exist relationship between P(Y) and Xj, b=0

>1 positive impact of Xj on P(Y), b>0

<1 negative impact of Xj on P(Y), b<0

From *odds* to P

$$odds = \frac{P}{1-P} = e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k}$$

$$P = \frac{odds}{1+odds} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k}}{1+e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k}} = \frac{1}{1+e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k)}}$$

Probability

If a subject of 22 years which buys on average 3 snacks per month and has seen on the tv the advertisement of the brand

$$Logit(P(purchase = 1)) = \quad 13{,}020 + 2{,}826(3) + 0{,}095(22) + 2{,}379(1)$$

$$P = \frac{1}{1 + e^{-(-13{,}020 + 2{,}826 \cdot 3 + 0{,}095 \cdot 22 + 2{,}379 \cdot 1)}} = 0.48$$

The probability of purchasing the snack of this brand is 48%