

# Factors that Cause University Students to Drop Out. An Alternative Modelling of Interaction Terms in Logistic Regression Models

*Determinanti dell'Abbandono Universitario. Un Modello Alternativo per i Termini d'Interazione nella Regressione Logistica*

Mariano Porcu<sup>1</sup>

Dip. Ric. Econ. e Soc. - Univ. di Cagliari  
mrporcu@unica.it

Vito M. R. Muggeo<sup>2</sup>

Dip. Sc. Statis. - Univ. di Palermo  
vito.muggeo@giustizia.it

**Riassunto:** Nei modelli di regressione logistica le interazioni tra variabili esplicative vengono solitamente valutate utilizzando il prodotto tra le variabili in questione. Tale procedura, comune agli usuali GLM, presenta alcuni svantaggi, primo fra tutti quello relativo all'ordine delle interazioni che viene mantenuto solitamente basso. In questo lavoro viene brevemente discusso un recente approccio che può essere utilizzato per analizzare le interazioni fra predittori. I risultati vengono illustrati per l'analisi dei fattori che possono favorire l'abbandono dagli studi per un campione di  $n = 573$  studenti universitari.

**Keywords:** University drop out, logit model, logic regression, Boolean logit.

## 1. Introduction

In the 'CATI' survey carried out at University of Cagliari in 2002, a sample of 573 students was selected and retrospectively followed up for ten years. The students belong to the cohort registered in the University for the a.y. 1991-92. At the time of survey, 161 of those students had already graduated, 122 were still enrolled at University and 290 had dropped out. Variables collected refer to the social demographic background of the students (sex, age at enrolment, residence, parents education) as well as to a set of 'structural' covariates (high school attended, exams passed, labour condition during university time, failures registered at high school, etc.). The goal of the study was to assess possible determinants of dropping out from University (Porcu and Puggioni, 2003).

The well-known logistic regression model is often used to model the probability of a specified event as a function of explanatory variables, including possible joint effects among covariates (e.g. Hosmer and Lemeshow, 1989). This task is usually performed fitting into the model the product among the variables themselves. Several reasons, technical (e.g. sparsity of the data) or theoretical (e.g. parsimony principle), usually force these interactions to be kept very simple (generally one-way or two-way) and furthermore the power of relevant tests is dramatically reduced. As a consequence, only main effects are usually included in spite of the joint effect that might turn out to be quite predictive or even useful in selecting groups (for instance in microarray-type data analysis).

In this paper the Authors briefly discuss a possible alternative approach to perform a regression analysis on the binary response  $Y$  considered: *dropping out* ( $Y = 1$ ) or *not-dropping out* ( $Y = 0$ ) from university. The method is described in section 2 and

---

<sup>1</sup>Indirizzo per corrispondenza: Mariano Porcu, Dipartimento di Ricerche Economiche e Sociali, Viale S. Ignazio 78, 09123 Cagliari. M. Porcu è autore dei §§1 e 3.

<sup>2</sup>Il lavoro è stato svolto con finanziamento 2002134337\_002. V. Muggeo è autore dei §§2 e 4.

section 3 includes the application on the ‘CATI’ data; finally section 4 is devoted to some discussions and conclusions.

## 2. A novel approach dealing with interactions in logit regression

Let  $Y$  be the binary response variable and  $X_0 \equiv 1, X_1, \dots, X_p$  the explanatory variables understood to explain variability in the response itself. The standard logistic regression model for the mean  $E[Y] = \pi$  is  $\text{logit}(\pi) = \sum_{j=0} \beta_j X_j = \mathbf{x}'\boldsymbol{\beta}$ .

Rather than including interactions by simple products among variables, in the so-called ‘Boolean logit regression’ (Braumoeller, 2003) we suppose that the binary response  $Y$  is thought to be produced by a *Boolean* or *logic* combination of some *conditions*  $\mathcal{A}_1, \dots, \mathcal{A}_k, \dots$ . For instance, given three conditions, theory could suggest that  $Y$  is produced when “ $\mathcal{A}_1$  and ( $\mathcal{A}_2$  or  $\mathcal{A}_3$ )” occur, namely  $Pr(Y = 1) = Pr(\mathcal{A}_1) \times Pr(\mathcal{A}_2 \cup \mathcal{A}_3)$ . Each condition, in turn, can occur with probability unknown but that can be modelled through a linear combination of explanatory variables on logistic scale,  $Pr(\mathcal{A}_k) = p_k = (1 + e^{-\eta_k})^{-1}$  such that  $\eta_k = \mathbf{x}'_k \boldsymbol{\beta}_k$  for every  $k = 1, 2, 3$ . Index  $k$  means that each condition can depend on its own explanatory variables through corresponding parameters; note that the same  $X_j$  may be included in different  $\eta_k$ , with no multicollinearity problem. Obviously, if the condition is only one, the boolean logit reduces to standard one.

Starting from the postulated model for  $\pi$ , using probability calculus it is possible to express  $\pi$  as a function of explanatory variables and relevant parameters through the  $p_k$ . For instance if it is assumed  $\pi = Pr(\mathcal{A}_1) \times Pr(\mathcal{A}_2)$  and  $\text{logit}(p_1) = \mathbf{x}'_1 \boldsymbol{\beta}_1$  and  $\text{logit}(p_2) = \mathbf{x}'_2 \boldsymbol{\beta}_2$ , the model and the likelihood  $Lik = \prod \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$  take the form:

$$\pi_i = p_{1i} \times p_{2i} \quad Lik(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^n (p_{1i} \times p_{2i})^{y_i} (1 - p_{1i} \times p_{2i})^{1-y_i} \quad (1)$$

Maximization of the log-likelihood and evaluation of its hessian matrix yields point estimates and relevant covariance matrix of the model parameters. So, once the occurrence of the event is explained in the language of complex causation, the consequent hypotheses can be expressed in terms of probability calculus (i.e., the combination of conditions  $\mathcal{A}_1, \dots, \mathcal{A}_k, \dots$ ) (Braumoeller, 2003).

A completely different approach is the so-called *logic regression* that attempts to determine a *Boolean* or *logic* combination of *binary* covariates as predictors (Ruczinski et al., 2003). That is, logic regression estimates the standard logit model with an additional (estimated) predictor  $L$ , expressed in a Boolean fashion depending on a specified set of binary explanatory variables,  $X_1, X_2, X_3$ , say. Although a particular algorithm has to be used for simultaneous estimation of the boolean predictor  $L$  and its parameters  $\delta$ , the logic regression is substantially an extension of classical logistic model.

In conclusion, it might be worthwhile to stress that the brackets are quite important since the Boolean statement “( $A$  and  $B$ ) or  $C$ ” is different from “ $A$  and ( $B$  or  $C$ )”.

## 3. Modelling the probability of dropping out

A logistic regression model has been fitted to our data modelling the response  $Y = 1$  if the student had dropped-out, and  $Y = 0$  if the student was still enrolled or graduated. The following dichotomous covariates have been considered, taking value ‘1=yes’: *at*

least one exam registered at first year (EXAM); ‘liceo’ attended as high school (the ‘liceo’ provides a classical education such as the one offered by old British ‘Grammar Schools’) (LICE; labour condition at university: worker (WORK); a failure registered at a year final examination at high school (FAIL); university studies did not begin soon after high school (AFTE); change of degree course (CHAN); residence not in the university city (RESI).

In addition to standard logit, a Boolean model has been fitted: here two conditions  $\mathcal{A}_1$ =‘lack of natural bent for studies’ and  $\mathcal{A}_2$ =‘difficulty/unease in relations with University’ have been assumed. In our opinion, the first condition is defined by a set of covariates that refer directly to students’ traits (EXAM, LICE, FAIL and WORK). On the other hand, the second condition is defined by the covariates related to objective circumstances that affect students’ performances, especially at the beginning of their university studies (AFTE, CHAN and RESI). The probability of students’ dropping-out,  $Pr(Y = 1) = \pi$ , is modelled as the interaction of  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , that is  $\pi = p_1 \times p_2$  where  $\mathcal{A}_1$  and  $\mathcal{A}_2$  are expressed as a function of explanatory variables, namely:  $\eta_1$ =EXAM+LICE+FAIL+WORK and  $\eta_2$ =AFTE+CHAN+RESI.

**Table 1:** Point estimates ( $\hat{\beta}$ ) and corresponding z values ( $z = \hat{\beta}/SE(\hat{\beta})$ ) for the two logit models.

| Covariates | standard      |          | Boolean       |          |
|------------|---------------|----------|---------------|----------|
|            | $\hat{\beta}$ | z -value | $\hat{\beta}$ | z -value |
| EXAM       | -1.942        | 8.91     | -2.739        | 5.63     |
| LICE       | -1.244        | 5.51     | -2.166        | 5.57     |
| WORK       | 0.464         | 1.19     | 1.228         | 1.77     |
| FAIL       | 1.832         | 4.06     | 2.301         | 2.66     |
| AFTE       | 1.783         | 2.31     | 1.451         | 2.08     |
| CHAN       | -1.109        | 2.86     | -1.741        | 3.66     |
| RESI       | 0.675         | 2.65     | 0.412         | 0.94     |
| log Lik    | -280.19       |          | -282.09       |          |

Table 1 shows parameter estimates for the standard logit model and the boolean one: for both models negative values mean a lower probability for ‘drop out’. Results are rather similar in terms of log likelihood and parameter estimates but the underlying models are different: in the former no interaction term shows significant effect on the response and therefore the model suggests that any variable influences the probability of dropping-out independently of the remaining ones. This result contrasts with the interpretation of the Boolean model where the response is produced by multi-vector interactions, namely interactions of vectors of variables. Thus, as it has been assumed that dropping-out from University depends jointly on  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , the parameters maintain tacitly an ‘interaction meaning’. Furthermore boolean logit allows to test for different roles of the same variables. For instance, the variable WORK could be included in both conditions: in  $\mathcal{A}_1$  it could measure people working since not really interested in university; in  $\mathcal{A}_2$ , such variable could select people which has to work to give financial support to own studies. A model with WORK also in  $\eta_2$  was fitted but the corresponding parameter was not significant and did not change estimates for the remaining coefficients.

As said, the interaction terms show no significant influence on the response in the standard logit model; 35% of students that drop out belong to just two covariate patterns (Christensen, 1990). These are referred to students that did or did not have their residence

in the university city, did not take any exam in their first year of university career, did not come from a 'liceo', did not change degree course, did not fail final year examinations at high school and began university studies soon after high school graduation. It is worthwhile to underline the negative effect on the probability of dropping out due to the parameter 'CHAN'. A change of university course at a glimpse could be considered a factor that increase the probability of a failure at university; rather it seems to be (for both models) a factor that reduces the risk of dropping out caused by a wrong choice of the field of study.

#### 4. Conclusions

In this paper we have tried to illustrate a possible alternative modelling for interactions in logit models. This method is quite recent, and as a consequence further research is needed to gain insight on possible its limits. For instance, in the boolean framework multi-vector interactions depending on any variable are fitted as well. On the other hand, the Boolean approach requires the underlying theory to be specified and it can be bothersome sometimes. Moreover, the specifying of the the postulated underlying factor could be seen as a very subjective judgement. Another drawback in the Boolean logit model concerns the interpretability of the parameters which are not log odds-ratios as they are in standard logistic or even logic regressions; this can be a problem when the researcher wants to summarise results by means of interpretable parameters. However, when the interest is to detect factors causing the (binary) response which interact one another, the boolean model allows to specify causal mechanisms that can be rather complex.

Whichever model is used (logic or Boolean regression), likelihood-based methods (e.g. the log likelihood itself or the AIC) can be employed to choose among different competing models.

#### References

- Braumoeller B.F. (2003) Causal complexity and the study of politics, *Political Analysis*, 11, 209–233.
- Christensen R. (1990) *Log-Linear Models*, Springer & Verlag, New York.
- Hosmer D.W. and Lemeshow S. (1989) *Applied Logistic Regression*, John Wiley & Sons, New York.
- Porcu M. and Puggioni G. (2003) Laurea ed abbandono universitario. Uno studio comparativo dei due eventi su un campione di immatricolati dell'Università di Cagliari, in: *Scoprire i rischi con l'analisi di segmentazione*, Fabbris L., ed., CLEUP, Padova, 25–40.
- Ruczinski I., Kooperberg C. and LeBlanc M. (2003) Logic Regression, *Journal of Computational and Graphical Statistics*, 12, 475–511.