BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LVIII No. 1/2006	31 - 36	Seria Matematică - Informatică - Fizică

Predicting students' status after the first session by using logistic regression

Cristian Marinoiu

Universitatea Petrol-Gaze din Ploieşti, Bd. Bucureşti 39, Ploieşti, Catedra de Informatică
e-mail: marinoiu_c@yahoo.com

Abstract

In this paper we build a regression logistic model which allows us to predict, with certain probability, students' status (if he/she has passed or not all the exams) in the informatics field, after the first winter session. This is made on the basis of the marks obtainedo in the entrance exam.

Keywords: *logistic regression, probability, estimate, bernoullian variable*

Introduction

After the winter session a student may have the following status:

- he/she has passed all the exams;
- he/she has failed at least one exam.

In order to obtain some facilities (reduction of the accommodation fees, scholarships for budgetary students) one of the compulsory conditions is to have passed all exams. Thus, a candidate to any faculty may be interested to predict his/her status after the winter session, on the basis of his/her entrance exam. Moreover, the staff of the faculties may wish to establish a minimal level of the entrance exams, in order to have a certain quality of their future students. The logistic regression model, presented in this paper, allows us to solve both problems stated above, for the freshmen in informatics field, of the Petroleum-Gas University of Ploiesti.

The logistic regression model

Let us consider the binary (bernoullian) variable Y and a random variable X. Since Y is a binary variable (*Y=1* or *Y=2*), then [1] $E(Y/x) = p$, $Var(Y/x) = p(1-p)$, where *x* is any value of X and $p = P(Y = 1/X = x)$.

We are interested in finding a relation between Y and X. Let us observe that if we tried to find a dependence of the form $E(Y/x) = \beta_0 + \beta_1 x$, we would have at least two inconvenient.

- The expected value $E(Y/x)$ must satisfy the relation

$$0 \le E(Y/x) \le 1 \tag{1}$$

- The variance $Var(Y/x) = p(1-p)$ is not constant; hence, an important assumption in regression model is violated.

The alternative is the logistic model

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \tag{2}$$

or, equivalently,

$$\log it(p) = \beta_0 + \beta_1 x, \tag{3}$$

where

$$\log it(p) = \ln\left(p/(1-p)\right), \tag{4}$$

The ratio $p/(1-p)$ is called the ratio of the odds and the above relation (4), the log*it* transformation or the log–odds.

Let us observe that the considered logistic model is a particular case (for *n=2*) of the following multinomial model [2]

$$P(Y = k / X = x) = \frac{\exp(\beta_{k0} + \beta_k^{'} x)}{1 + \sum_{i=1}^{n-1} \exp(\beta_{io} + \beta^{'}{}_i x)}, \quad k = 1,2,\ldots, n-1, \tag{5}$$

$$P(Y = n / X = x) = \frac{1}{1 + \sum_{i=1}^{n-1} \exp(\beta_{io} + \beta_i^{'} x)} \tag{6}$$

or, equivalently

$$\ln \frac{P(Y = k / X = x)}{P(Y = n / X = x)} = \beta_{k0} + \beta_k^{'} x, \qquad k = 1,\ 2,\ldots, n-1 \tag{7}$$

Predicting the students' status after the winter session

In table 1 we present the entrance exam scores and the students' status after the winter session. We have adopted the following conventions:

- Y=1, if the student has passed all the exams
- Y=2, if the student has failed at least one exam

Table 1. Entrance exams marks and students' status after the first winter session

Student number	**Marks** *ma*	**Student status**	**Student Number**	**Marks** *ma*	**Student Status**
1	6.53	2	24	8.29	2
2	6.59	2	25	8.36	2
3	6.92	2	26	8.41	2
4	6.94	2	27	8.47	2
5	6.96	2	28	8.47	2
6	7.12	2	29	8.57	2
7	7.36	2	30	8.63	2
8	7.54	2	31	8.65	1
9	7.54	2	32	8.66	2
10	7.58	2	33	8.68	1
11	7.74	2	34	8.69	1
12	7.84	1	35	8.76	2
13	7.93	2	36	8.86	2
14	7.97	2	37	8.88	2
15	8.00	2	38	8.91	1
16	8.05	2	39	8.97	1
17	8.10	1	40	9.18	2
18	8.14	2	41	9.22	1
19	8.19	2	42	9.30	1
20	8.21	2	43	9.31	1
21	8.25	2	44	9.68	1
22	8.28	1	45	9.93	1
23	8.28	2			

In order to know if a model of the form (2) can explain the relation between Y and X, we will reorganize the data from the table 1 as follows:

- regroup the marks of the entrance exam into 7 intervals; choose the midpoint of each interval as a value of the independent variable *ma*
- for each interval *i* find the number of students labeled by 1 and calculate the observed probability p_i of *Y=1* for each interval x_i
- in order to be able to calculate $\log it(p_i)$ one should make the following adjustments: replace the probability p_i=0.0 with p_i=0.01 and the probability p_i=1.0 with p_i=0.99

As a result of our reorganization we have obtained the data presented in the table 2.

Table 2. Training data for the logistic regression model

Interval	[6.50-7)	[7, 7.50)	[7.50-8)	[8.00-8.50)	[8.50, 9)	[9, 9.50)	[9.50-10]
ma_i	6.75	7.25	7.75	8.25	8.75	9.25	9.75
p_i	0.01	0.01	0.14	0.21	0.45	0.75	0.99
$\log it(p_i)$	-4.6	-4.6	-1.82	-1.32	-20	1.10	4.60

Now we can apply the standard regression methodology for the linear regression model

$$\log it(p_i) = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1,2,...,n \qquad (8)$$

where:

β_0, β_1 are the unknown parameters;

p_i, ma_i, $\log it(p_i)$ are presented in table 2;

ε is the additive error.

Using SPSS [3] we found that the value of the coefficient of determination is 0.934. This value shows us there is a strong linear dependence between the dependent variable $\log it(p_i)$ and the independent variable X.

In order to estimate the parameters β_0 and β_1 we can use the least squared method or the maximum likelihood method.

The least squared approach

The least squares estimations of the coefficients for the model are obtained using SPSS again. They are: $\hat{\beta}_0 = -24.89,\ \hat{\beta}_1 = 2.90$.

Below, in figure (1) we present the plot and the linear fit.

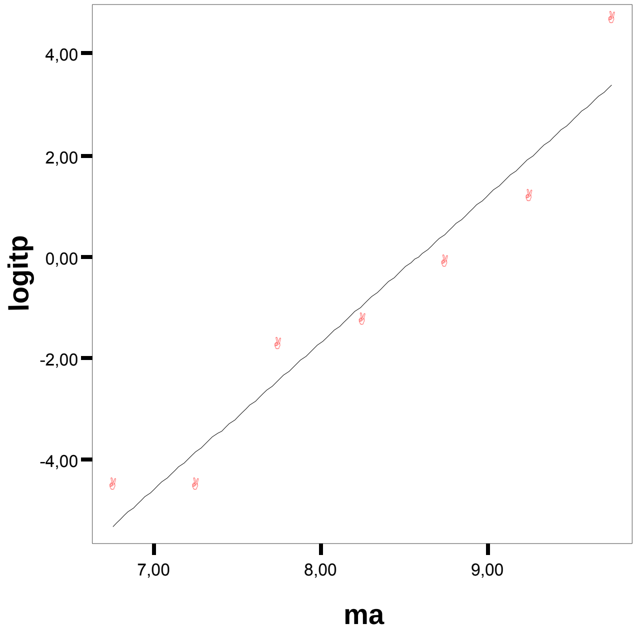

Fig. 1. Ordinary linear regression

From the relation

$$\log it(\hat{p}) = \hat{\beta}_0 + \hat{\beta}_1 * ma \quad , \tag{9}$$

we can find immediately the desired logistic function as

$$\hat{p} = \frac{e^{-24.89\ \ +2.90_1 * ma}}{1 + e^{-24.89+2.90\ \ * ma}} \tag{10}$$

The maximum likelihood approach

Unfortunately, due to the non normality of the errors model, the least squared method cannot allow us to make inferences about parameters. For this reason, we usually fit by maximum likelihood, or for convenient calculus, by log – likelihood method. Because the maximum likelihood estimators are asymptotically normally distributed, this allows us to compute confidence intervals and perform statistical tests. The results obtained using Matlab[4] are the following (table 3):

Table 3. Maximum Likelihood Estimates

Coefficient	Estimates	Standard error	t-test	p-value
β_0	$\hat{\beta}_0 = -19.8393$	6.7787	-2.9267	0.0034
β_1	$\hat{\beta}_1 = 2.2593$	2.8511	2.8511	0.0044

One can observe that both interested *p-values* (0.0034 and 0.0044) are less than the significance level $\alpha = 0.05$. Therefore, both the intercept ($\hat{\beta}_0$) and the slope ($\hat{\beta}_1$) coefficients are significantly different from zero. This shows that the entrance exam mark (*ma*) is a significant predictor for the students' status after the first winter session. Hence, the obtained logistic function is as follows:

$$\hat{p} = \frac{e^{-19.8393+2.2593*ma}}{1+e^{-19.8393+2.2593*ma}} \tag{11}$$

The logistic regression $\hat{p}$ graphed against the data is obtained using also Maltab software The s-shaped graph is presented below, in figure 2.

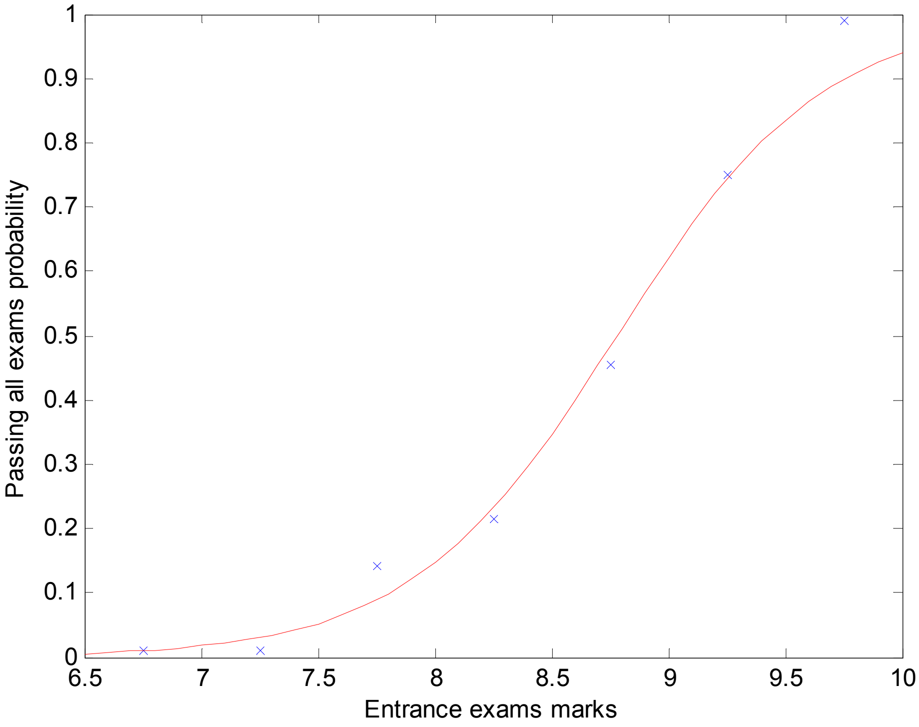

Fig. 2. The logistic regression model

Conclusions

The logistic regression models built in this paper show the dependence between the probability of passing all exams in the winter session and the entrance exam marks; this is done for the freshmen in the informatics field from Petroleum-Gas University of Ploiesti. Using SPSS and Matlab we have obtained the least squared estimators and the maximum likelihood estimators of the unknown parameters. Further, in the last case, the significance of the coefficients of the model has been established and the s-shaped graph of the logistic model has been presented.

The obtained models constitute practical instruments to estimate the performance level of the future students, For example, by inverting the relation (11) we find that, in order to have a percent of at least 50% students without failed exams, the minimum mark in the entrance exam must be 8.78.

References

1. Klimov, G. - *Probability Theory and Mathematical Statistics,* Mir Publishers, Moscow, 1986
2. Hastie, T., Tibshirani,G., Friedman J. - *The elements of statistical learning, Data Mining, Inference and Prediction,* Springer-Verlag, New York, 2001
3. * * * - SPSS for Windows, version 11
4. * * * - Matlab 7.0., The language of technical computing

Predicția statutului studenților după prima sesiune utilizând regresia logistică

Rezumat

În acest articol construim un model de regresie logistică; modelul ne permite să prevedem cu o anume probabilitate, statutul unui student din domeniul informatica (dacă este integralist sau nu) după prima sesiune de iarnă. Predicția se bazează pe notele obținute la concursul de admitere.