Analysis between GPA and TOEFL Score For Postgraduate Student of Sports Education using Bivariate Logistic Regression

A Sofro†, A Oktaviarina†, E Mintarto ‡ and O Wiriawan‡

† Mathematics Department, Universitas Negeri Surabaya, East Java, Indonesia
‡ Sport Science Faculty, Universitas Negeri Surabaya, East Java, Indonesia
E-mail: ayuninsofro@unesa.ac.id

Abstract. Due to the globalization era, there is the challenge of competition in all fields including finding work in sports education. To meet the stakeholder’s needs, it is necessary to have some criteria for prospective workers and also a well-planned process in academic. Grade Point Average (GPA), is one of the most important achievements during the postgraduate programme. Meanwhile, another academic component is English as a Foreign Language Test Score (TOEFL) which also has a very significant impact on seeing the job.

This paper will analyze valuable factors that influence GPA and TOEFL using a bivariate approach. In this case, the GPA will be defined as two levels, i.e low and high. TOEFL is also able to define as two categories, i.e more and less than 450. Hence, bivariate logistic regression is the most appropriate approach. The result shows that the coefficient of age has strong evidence to reject the null hypothesis at a significant level of 0.05. It means that the variable of age has a huge contribution to influence the score of the TOEFL for the postgraduate programme in sports education.

1. Introduction

GPA is a reflection of the level of academic success of students, which contains the personal nature of students such as discipline, perseverance in learning, hard work and mindset development. GPA is one indicator that is considered by a company or institution in recruiting employees. There are also other factors such as aptitude tests, interviews, and work experience. Another academic component is TOEFL which is considered by companies for recruiting employees.

Since the last decade, Indonesia has launched the era of globalization and free markets. This era requires expertise to communicate internationally. The competition is in all fields including for seeking a job. To meet the criteria of the shareholders, qualified human resources is needed. From an academic point of view, a fairly long and planned process is needed to realize the criteria for prospective workers.

Sports education in Universitas Negeri Surabaya is the only postgraduate programme in East Java. The goal is to produce graduates who are of high quality and are capable of carrying out
development tasks in the field of sports for the community. The postgraduate students in sports education come from various circumstances, i.e. original, high school status. To provide a high quality of postgraduate students in the programme, further investigation needs to be done such as analysis for GPA and the score TOEFL as basic information for making decisions.

The study between score TOEFL and GPA has been developing massively. Although the TOEFL scores cannot be used to predict the academic performance, the student who gets higher TOEFL iBT tend to earn a higher GPA as academic performance, see [1] and [2]. Meanwhile, some evidence shows that there is a strong correlation between TOEFL iBT and GPA which can be found in [3].

The GPA can be categorized as high if it exceeds 3.5 and low for others. Similar to the previous measurement, the TOEFL score, it will be high if it is above 450 and low for others. Therefore, the GPA and TOEFL scores follow the binomial bivariate distribution. One method for finding relationships where response variables have bivariate categorically (high and low), with factors of one or more independent variables is using bivariate binomial regression.

The logistic regression is very well known method to accommodate categorized response, see [4], [5] and [6]. Meanwhile, the application of logistic regression in education also can be found in [7] and in public health see [8] and [9]. Many data including education become more than one response variable and complex. This issue has been accommodated by some developing logistic regression theories, say bivariate approach see [10]. The application of the epidemiology model can also be seen in [11]. The approach has been developed under VGAM package for user friendly, see [12], [13] and [14]. Finally, the package has been improved in 2008 by Zelig package, see [15] and [16].

In this paper, we will focus on analyzing the GPA and the score of TOEFL for Postgraduate students of sports education using bivariate logistic regression. The result of the models will be also discussed.

2. Method

In this section, to understand a bivariate logistic regression, we start with a simple approach. i.e logistic regression.

2.1. Logistic Regression Model

Suppose that $y_i$ is categorical response variable with the value of either 0 or 1 and $i = 1, ..., n$. We can define with value 1 if it is success and 0 for failure. The covariate variables are defined as $x_i$, where also $i = 1, ..., n$ and $n$ is the sample size. The model is following a binomial distribution with logit as the link function. Hence, the model of logistic bivariate is following

$$y_i | \pi_i \sim Bin(1, \pi_i),$$

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = x_i^T \beta.$$  \hspace{1cm} (1)

The logit function can be rewritten down as following

$$\frac{\pi_i}{1 - \pi_i} = \exp(x_i^T \beta) \hspace{1cm} (2)$$

$$\pi_i + \pi_i \exp(x_i^T \beta) = \exp(x_i^T \beta) \hspace{1cm}$$

$$= \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}.$$  \hspace{1cm} (3)

Since each $y_i$ is defined as a binomial distribution, hence the probability of $y_i$ can be represented as follows

$$P(y_i | \beta) = \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}, \hspace{0.5cm} i = 1, ..., n.$$
In order to estimate the parameters ($\beta$), we will use maximum likelihood estimator approach. The first procedure is to construct the likelihood function of the logistic regression model. The likelihood function can be written down as follows

$$L(\beta|y_i)) = \prod_{i=1}^{n} P(y_i|\beta)$$

$$= \prod_{i=1}^{n} (\pi_i^{y_i} (1 - \pi_i)^{1-y_i})$$

$$= \prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} (1 - \pi_i)$$

$$= \prod_{i=1}^{n} \left(\exp(x_i^T \beta)^{y_i} \left(1 - \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right)\right). \quad (4)$$

Then, we involve a natural log to the likelihood function for convenient, hence the log likelihood function can be defined as follows

$$l(\beta) = \sum_{i=1}^{n} \left( y_i(x_i^T \beta) - \log(1 + \exp(x_i^T \beta))\right). \quad (5)$$

Finally, we maximize the equation (5) to obtain the parameter model. One of the method to optimize is using Newton-Raphson method.

2.2. Bivariate Logistic Regression

In this model, there are two categorical (binary) response variables which can be defined as $y_1$ and $y_2$. To construct the joint probability of the model, we can start from the marginal of dependent variables and the odds ratio which measures the association between the two categorical response variables. We also define as well that $y_{vw}$ are the random bivariate variable which can contain two values, either 1 or 0. Hence, in this case we will involve four random bivariate variables, i.e $y_{11}, y_{10}, y_{01}$ and $y_{00}$. The probability of the random bivariate variables are $\pi_{11}, \pi_{10}, \pi_{01}$ and $\pi_{00}$ respectively. Thus, the model with the logit link function can be defined as follows

$$y_{11} \sim Bin(y_{11}|\pi_{11}),$$
$$y_{10} \sim Bin(y_{10}|\pi_{10}),$$
$$y_{01} \sim Bin(y_{01}|\pi_{01}). \quad (6)$$

and the odd ratio as the association measurement of two dependent variables are defined as

$$\Psi = \frac{\pi_{00}\pi_{01}}{\pi_{10}\pi_{11}} \quad (7)$$

where

$$\pi_{11} = P(y_1 = 1, y_2 = 1)$$
$$\pi_{10} = P(y_1 = 1, y_2 = 0)$$
$$\pi_{01} = P(y_1 = 0, y_2 = 1)$$
$$\pi_{00} = P(y_1 = 0, y_2 = 0).$$
Meanwhile the logit link function for marginal probability for each dependent variables are follows

\[
\log \left( \frac{\pi_j}{1 - \pi_j} \right) = \eta_j, \quad j = 1, 2 \\
\log(\Psi(x)) = \eta_3
\]

where \( \eta_j = \beta_j^T x \). The marginal probability of binary response variables with \( p \) independent variables can be rewritten as following

\[
\pi_1 = P(y = 1) = \frac{\exp(\beta_0 + \beta_{11} x_1 + \beta_{21} x_2 + \ldots + \beta_{p1} x_p)}{1 + (\exp(\beta_0 + \beta_{11} x_1 + \beta_{21} x_2 + \ldots + \beta_{p1} x_p))} \\
\pi_2 = P(y = 2) = \frac{\exp(\beta_0 + \beta_{12} x_1 + \beta_{22} x_2 + \ldots + \beta_{p2} x_p)}{1 + (\exp(\beta_0 + \beta_{12} x_1 + \beta_{22} x_2 + \ldots + \beta_{p2} x_p))} \\
\Psi = \exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_p x_p)
\]

The other quantities can be calculated using formula as following

\[
\pi_{11} = \left\{ \begin{array}{ll} \frac{1}{2}(\Psi - 1)^{-1} - a - \sqrt{a^2 + b} & ; \Psi \neq 1 \\ \pi_1 \pi_2 & ; \Psi = 1 \end{array} \right. \\
\pi_{10} = \pi_1 - \pi_{11} \\
\pi_{01} = \pi_2 - \pi_{11} \\
\pi_{00} = 1 - \pi_{10} - \pi_{01} - \pi_{11}
\]

where \( a = 1 + (\pi_1 + \pi_2)(\Psi - 1) \) and \( b = -4\Psi(\Psi - 1)\pi_1 \pi_2 \).

Let the two binary bivariate response variables are denoted as \((y_{1i}, y_{2i})\) with the values either 1 or 0, where \( i = 1, \ldots, n \), \( n \) is the number of observation. We have discussed in the beginning sub section, we can recall that the model has defined some binary bivariate random variables independently as , i.e \( y_{1i}, y_{0i}, y_{01} \) and \( y_{00} \) and the probability are \( \pi_{11}, \pi_{10}, \pi_{01} \) and \( \pi_{00} \) respectively. Hence, the probability density function of the model can be written as follows

\[
P(y) = \begin{cases} 
\pi_{vuw}^{y_{vuw}}; & v = 0, 1 \quad w = 0, 1, \\
\pi_{11}^{y_{11}} \pi_{10}^{y_{10}} \pi_{01}^{y_{01}} \pi_{00}^{y_{00}}; & \end{cases}
\]

Then the likelihood function of the model can be denoted as

\[
L(\beta) = \prod P(y) = \prod \pi_{vuw}^{y_{vuw}} \pi_{11}^{y_{11}} \pi_{10}^{y_{10}} \pi_{01}^{y_{01}} \pi_{00}^{y_{00}}.
\]

The log likelihood function of equation 9 are written down as

\[
l(\beta) = \sum (y_{11} \ln \pi_{11} + y_{10} \ln \pi_{10} + y_{01} \ln \pi_{01} + y_{11} \ln \pi_{11}).
\]

We need to maximize the equation 10 to estimate the parameter of model \((\beta)\).

2.3. Data and Variables

In this paper, we have surveyed postgraduate students in sports education programme, Universitas Negeri Surabaya in 2018 using a random sampling technique. At the end of the survey, we have got 31 participants. Some of the dependent variables are GPA (\(y_1\)) and TOEFL score (\(y_2\)), whilst independent variables are GPA and TOEFL score when an undergraduate programme, age, and employment status.
### 3. Result and Discussion

Table 1 shows the statistics descriptive for some variables, i.e. GPA and TOEFL scores. Meanwhile, the statistics descriptive of the independent variables such as age and employee status is provided in Table 2.

| Variable             | Max  | Min  | Mean |
|----------------------|------|------|------|
| GPA Undergraduate    | 3.5  | 2.98 | 3.32 |
| GPA Postgraduate     | 3.88 | 3.28 | 3.72 |
| TOEFL Score (Undergraduate) | 470 | 384 | 412.93 |
| TOEFL Score (Postgraduate) | 507 | 333 | 430.12 |

Table 1. Descriptive statistics for some variables

In this paper, we start to analyze the data using the simple approach, i.e logistic regression model. In this case, the dependent variable is defined as the TOEFL scores in the postgraduate programme, whilst the independent variables are GPA and TOEFL scores in the undergraduate program, age, and employee status. The estimated parameter using the approach can be found in Table 3. From Table 3 shows that the coefficient of the age variable has strong evidence to reject the null hypothesis at $\alpha = 0.05$. It means that the variable is a significant factor to influence the TOEFL scores. Meanwhile, we also can express that the coefficient for age is $-2.280$ which refers to the log odd ratio between age (20-25) group and age (25-30) group since we denote that the younger group is null. The odd ratio equals $0.1022842$ which means that the odd for age (25-30) is about 10% higher than the odd for age (20-25) to have a good TOEFL score. The model using logistic regression can be written down as follows

$$
\pi = \frac{\exp(0.300 + 0.384x_1 + 0.059x_2 + 1.212x_3 - 2.280x_4)}{1 + \exp(0.300 + 0.384x_1 + 0.059x_2 + 1.212x_3 - 2.280x_4)}.
$$

Based on the [2], a student who earns a higher TOEFL score tend to have a good GPA as well. Then, we provide analyzing the GPA and TOEFL score simultaneously using the bivariate approach. The two dependent variables are GPA and TOEFL score in the postgraduate programme, whilst the GPA and TOEFL in undergraduate, employee status and age are independent variables. The estimated parameters using bivariate logistic regression can be found in Table 4.

| Coefficients       | Estimate | Std Error | Z value | P-value |
|--------------------|----------|-----------|---------|---------|
| (Intercept)        | 0.300    | 4.054     | 0.074   | 0.941   |
| GPA undergraduate  | 0.384    | 0.475     | 0.809   | 0.418   |
| TOEFL undergraduate| 0.059    | 0.520     | 0.115   | 0.908   |
| Employee status    | 1.212    | 0.983     | 1.233   | 0.217   |
| Age                | -2.280   | 0.938     | -2.430  | 0.015   |

Table 3. Estimated parameters TOEFL data using Logistic Regression Model

Table 2. Descriptive statistics for some independent variables

| Variable         | Category      | Frequency | Percentage (%) |
|------------------|---------------|-----------|----------------|
| Age              | 20-25         | 13        | 41.9           |
|                  | 25-30         | 18        | 58.1           |
| Employee Status  | Employee      | 19        | 61.3           |
|                  | Unemployee    | 12        | 38.7           |

Table 2. Descriptive statistics for some independent variables
Coefficients | Estimate | Std Error | Z value | P-value
--- | --- | --- | --- | ---
(Intercept):1 | 2.976 | 3.853 | 0.773 | 0.439
(Intercept):2 | -0.572 | 3.887 | -0.147 | 0.882
(Intercept):3 | 1.627 | 0.990 | 1.643 | 0.100
GPA undergraduate :1 | -0.073 | 0.418 | -0.176 | 0.860
GPA undergraduate :2 | 0.417 | 0.460 | 0.908 | 0.364
TOEFL undergraduate:1 | -0.459 | 0.512 | -0.897 | 0.369
TOEFL undergraduate:2 | 0.127 | 0.496 | 0.258 | 0.796
Employee status :1 | 0.620 | 0.874 | 0.709 | 0.478
Employee status :2 | 1.203 | 0.951 | 1.265 | 0.206
Age :1 | -0.361 | 0.655 | -0.551 | 0.581
Age :2 | -1.962 | 0.877 | -2.235 | 0.025

Table 4. Estimated parameters using the Bivariate Logistic Regression

From Table 4, we can see that the coefficient of age variable for logit function in the second response variable (TOEFL scores) has strong evidence to reject the null hypothesis at $\alpha = 0.05$. It means that the age variable is a significant aspect to contribute to the TOEFL score which is similar to the logistic regression approach. From Table 4, we also can express that the coefficient of age for the second response variable is $-1.962$ referring to the log odd ratio between age (20-25) group and age (25-30) group. It means that the odd of age (25-30) group is 14% higher that odd of age (20-25) group to earn a high TOEFL score due to the older group signed as 1. The model using the bivariate approach from Table 4 can be written down as follows

$$
\pi_1 = \frac{\exp(2.976 - 0.073x_1 - 0.459x_2 + 0.620x_3 - 0.361x_4)}{1 + \exp(2.976 - 0.073x_1 - 0.459x_2 + 0.620x_3 - 0.361x_4)}
$$

$$
\pi_2 = \frac{\exp(-0.572 + 0.417x_10.127x_2 + 1.203x_3 - 1.962x_4)}{1 + \exp(-0.572 + 0.417x_10.127x_2 + 1.203x_3 - 1.962x_4)}
$$

and the odd ratio is $\Psi = \exp(1.627) = 5.0923$.

For the goodness of fit model, we provide Akaike Criterion Information (AIC) to evaluate models in Table 5. From the Table 5 shows that bivariate approach provides the smallest AIC

| Model                     | AIC     |
|---------------------------|---------|
| Logistic Regression       | 41.76   |
| Bivariate Logistic Regression | 40.79   |

Table 5. AIC value for two models

values. It can be said that the bivariate logistic model is the best model for analyzing GPA and TOEFL scores.

4. Conclusion
Age is a significant variable to influence TOEFL scores using univariate and bivariate logistic regression. The bivariate approach has provided the best model for the data as producing the smallest value of AIC.

References
[1] Cho Y and Bridgeman B 2012 Relationship of TOEFL iBT Scores to Academic Performance: Some Evidence from American Universities *Language Testing* vol. 29 (3) pp 421-442
[2] Vu L T and Vu P H 2013 Is the TOEFL Score A reliable Indicator of International Graduate Student’s Academic Achievement in American Higher Education? *International Journal on Studies in English Language and Literature* vol. 1

[3] Ginther A and Yan X 2017 Interpreting the relationship between TOEFL iBT Scores and GPA: Language Proficiency, Policy and Profiles *Language Testing* vol. 35(2) pp 271-295

[4] Berger D E 2017 Introduction to Binary Logistic Regression and Propensity Score Analysis *Working paper*

[5] Tranmer M and Elliot M Binary Logistic Regression *Canttie Marsh Centre for Census and Survey Research*

[6] Sperandei S 2014 Understanding Logistic Regression Analysis *Biochemia Medica* vol. 24 (1) pp 12-18

[7] Peng C Y J, Lee K L and Ingersoll G M 2002 An Introduction to Logistic Regression Analysis and Reporting *The Journal of Education Research* vol. 96 pp 3-14

[8] Park H A 2013 An Introduction to Logistic Regression : From Basic concept to interpretation with particular Attention to Nursing Domain *Journal Korean Academic Nursing* vol. 43 pp 164 - 164

[9] Raspriranty D I and Sofro A 2018 Analysis of Hypertension Disease using Logistic and Probit Regression *Journal of Physics Conference Series* vol. 1108(1) : 012054

[10] McCullagh P and Nelder J A 1983 Generalized Linear Models, Chapman and Hall

[11] Fitzmaunce G M, Laird N M, Zahner G E P, and Daskalakis C 1995 Bivariate Logistic Regression of Childhood Psychopathology Ratings Using Multiple Informants *American Journal of Epidemiology* vol. 142 pp 1194 -1203

[12] Yee T W The VGAM Package for Categorical Data Analysis

[13] Yee T W and Hastie T J 2003 Reduced Rank Vector Generalized Linear Models *Statistical Modelling* col. 3 pp 15 -41

[14] Yee T W 2008 VGAM family functions for Bivariate Binomial Responses

[15] Imai K, King G, and Lau O 2008 Zelig: Everyone’s Statistical Software

[16] Imai K, King G and Lau O 2008 Toward a common Framework for Statistical Analysis and Development *Journal of computational and Graphical Statistics* vol. 17 pp 892 - 913