Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666
Vol. 4, Issue 2 (2011): 105-117

APPLYING LOGISTIC REGRESSION MODEL TO THE EXAMINATION RESULTS DATA

Goutam Saha
Govt. Degree College, Kamalpur, Tripura
E Mail: goutam.stat@gmail.com

Abstract

The binary logistic regression model is used to analyze the school examination results (scores) of 1002 students. The analysis is performed on the basis of the independent variables viz. gender, medium of instruction, type of schools, category of schools, board of examinations and location of schools, where scores or marks are assumed to be dependent variables. The odds ratio analysis compares the scores obtained in two examinations viz. matriculation and higher secondary.

Key Words: Logistic Regression Model, Correspondence Analysis, Odds Ratio, Likelihood Ratio Tests, Wald Statistic, Shortest Confidence Interval.

1. Introduction

Neither two students' nor two schools are identical. Students' differ in gender, culture, religion, language, home environment, financial status of parents etc., whereas the schools differ in size of students, quality of teacher, infrastructure, location of the school, aid provided by the government etc.. Obviously performance of the students measured in terms of scores or grades obtained by them in examinations varies from student to student and school to school. The variability in scores is a function of social climate which has to be studied and analyzed scientifically. The history of analyzing the students performance is as old as history of education. However formal presentation of analysis started around early thirties of the 20^{th} century.

The performance measure corresponding to different independent variables may be analyzed using logistic regression analysis. Logistic regression has been successfully employed in social science, biostatistics, genetics and demographic issues, but as far as school examinations are concerned, not many research articles are available.

This paper deals with presentation and analysis of examination results of Tripura: north-east India. The matriculation results of the students admitted in higher secondary (science stream) are collected, along with their higher secondary scores. The data collected consist of grades of the students in both the examinations along with their scores in English, Mathematics and Science subjects.

It is assumed that scores of students are affected by social environment controlled by the parameters viz. (i) gender (male, female), (ii) medium of instruction (English, Bengali), (iii) type of schools (boys', girls', and co-educational), (iv) category of schools (Govt., non-Govt.), (v) board of examinations i.e., Tripura Board of Secondary Education (TBSE), Central Board of Secondary Education (CBSE) and Indian Certificate of School Examination (ICSE) and (vi) location of schools (urban,

rural). The logistic regression approach has been adopted to study the examination scores under the variables mentioned above.

Scores of students are partitioned into two sets viz. [0, 45) and [45, 100]. Since in the above mentioned examination 45% and above (obviously 60% and above means first class) marks indicate second class, hence the students are classified as belonging to two different categories as far as their scores are concerned. As a consequence idea of binary logistic regression analysis seems to be appropriate when scores are functions of independent variables mentioned above.

A brief review of literature on this subject is included in section 2. Section 3 has data source. Materials and methods are shown on section 4. Section 5 contains numerical analysis of binary logistic regression. Results and interpretations are incorporated in section 6.

2. Review of Literature

Peng et al (2002) applied the logistic regression technique to compare the sample data of gender and recommendation for remedial reading instruction. David et. al. (2001) used logistic regression analysis to determine whether grade point average and hours of education is significant predictor of performance on the national athletic trainers' association board of certificate examination. E. L. Dey and Astin A. W. (1993) studied the focus on the practical implications of applying logistic regression, probit analysis and linear regression to the problem of predicting the college student retention. Jason et. al. (2001) analyzed the logistic regression method to predict the probability of passing a course based on the scores on California chemistry diagnostic test at two different institutions with two different instructors over multiple years. Robert B. and Vaughan B. (2006) checked which factors were key in enabling or constraining a students' ability to close the achievement gap during the school results. Erin et al. (2010) used multilevel logistic regression analyses, to explore the school and student level characteristics associated with moderate and high levels of physical activity among school students. In series of articles Sarma and Sarmah (1999), Saha and Sarmah (2010) and Saha and Sarmah (2011) discussed the probabilistic analysis and testing of some important hypothesis using Markov chain.

3. Data Source

1002 samples have been collected from the entire state of Tripura – a part of north-east India, through a Minor Research Project (MRP) entitled "Prospects and Problems of Educational development (Higher Secondary Stage) in Tripura - An in-depth Study" sponsored by University Grants Commission (UGC), New Delhi, India.

In this survey from a total population of 225 schools, offering both matriculation and higher secondary (science stream) courses, a sample of 75 schools are randomly selected and data related to examination scores are collected for analysis. The results of the students in both the examinations are assumed to be influenced by the variables gender, medium of instructions, type of schools, category of schools, board of examinations and location of schools.

4. Materials and Methods

The Model

Let us consider the indicator variable $Y_i = 1$, if $Y_i \in [45, 100]$

$= 0$, if $Y_i \in [0, 45)$

Now let us assume $P_r(Y_i=1) = \theta = 1 - P_r(Y_i=0)$

Where, θ can be written as,

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}$$

Where as, α = the constant of the equation and β = the coefficient of the predictor variables X_i for i = 1, 2, …, n.

An alternative form of the logistic regression equation is:

$$\mathbf{logit}\left[\theta(\mathbf{x})\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i \quad (1)$$

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable.

Logistic regression calculates the probability of success over the probability of failure; the results of the analysis are in the form of an odds ratio. The odds ratio is a measure of effect size (Westergren A. 2001), describing the strength of association or non-independence between two binary data values. It is used as a descriptive statistics, and plays an important role in 'logistic regression'. Unlike other measures of association for paired binary data such as the relative scores, the odds ratio treats the two variable being compared symmetrically, and can be estimated using some type of non-random samples.

Hypothesis testing in logistic regression involves reasoning by contradiction. The first assumption or the null hypothesis is that, the predictor coefficient is zero in the population. Hypothesis test tell whether there is sufficient evidence in the sample data to reject the null hypothesis and therefore to accept the alternative hypothesis that the predictor variable coefficient differ from zero. Confidence intervals can be used for hypothesis testing as well as for regression coefficients.

The odds ratio may be presented as

$$L_i = \ln\left[\frac{P_i}{1 - P_i}\right] \tag{2}$$

With $P_i = P_r[X_i = 1] = 1 - P_r[X_i = 0]$, where X_i is the independent variable corresponding to i^{th} category for i = 1, 2, …, n.

5. Numerical Analysis of Binary Logistic Regression

In this paper binary logistic regression analysis is performed with dependent variable of total marks obtained in higher secondary and matriculation examination, in presence of independent variables gender, medium of instruction, type of schools, board of examinations, category of schools and location of schools.

By applying the method of Correspondence Analysis (CA), it is observed that, there exists a significant correlation between (i) type of schools and board of examinations with r = 0.228, (ii) type of schools and category of schools with r = 0.141 and (iii) category of schools and board of examinations with r = 0.266. Here, we perform binary logistic regression analysis.

6. Results and Interpretation

Data Variable	Data Explanation	Data Type	Conditioned Used
	Dependent	**Variable**	
Marks in higher secondary / matriculation examination	Not-Satisfied Satisfied	Binary	0 - Not Satisfied 1 - Satisfied
	Independent	**Variables**	
Gender	Gender of Student	Binary	0 - Female 1 – Male
Medium	Medium of Instruction	Binary	0 – English 1 - Bengali
School Type	Type of Schools	Categorical	0 – Boys' 1 – Girls' 2 – Co-Educational
Board	Board of Examinations	Categorical	0 – TBSE 1 – CBSE 2 - ICSE
Schools Category	Category of Schools	Binary	0 – Govt. 1 – Non-Govt.
Location	Location of School	Binary	0 – Urban 1 - Rural

Table 1: Coding of Variables affecting Higher Secondary and Matriculation Results

Variable	Category	Frequency	%
Gender	Female	580	42.10
	(Male)	422	57.90
Medium	English	321	32.00
	(Bengali)	681	68.00
School Type	Girls'	204	20.40
	Boys'	214	21.40
	(Co-Education)	584	58.30
Board	TBSE	707	70.60
	CBSE	151	15.10
	(ICSE)	144	14.40
School Category	Govt.	432	43.10
	(Non-Govt.)	570	56.90
Location	Urban	476	47.50
	(Rural)	526	52.50

Table 2: Categorical Variables Affecting Higher Secondary and Matriculation Results

Here the reference category is shown in parenthesis (2nd column in Table 2) whose odds ratio is '1'.

Classification	Frequency	Percentage (%)
Not-Satisfied	135	13.50
Satisfied	867	86.50
Total	1002	100.00

Table 3: Classification of Total Marks in Higher Secondary Examination

Classification	Frequency	Percentage (%)
Not-Satisfied	053	05.30
Satisfied	949	94.70
Total	1002	100.00

Table 4: Classification of Total Marks in Matriculation Examination

By dropping one of the variables which are correlated, i.e., type of schools, board of examinations and category of schools are not used simultaneously as independent variables. Hence, we perform binary logistic regression analysis in three stages. In the second column of the following tables 'B' represents the coefficient for the constant (also called the 'intercept') in the null model.

6.1 Analysis (Stage – I)

Here we consider higher secondary result as a dependent variable; where as gender, medium of instruction, school type and location of schools are independent variables.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender (Female)	1.008	0.303	11.098	1	0.001	2.740	1.514	4.959
Medium (English)	0.808	0.243	11.089	1	0.001	2.244	1.395	3.611
School_Type			8.421	2	0.015			
School_Type (Girls)	-0.402	0.348	1.333	1	0.248	0.669	0.338	1.324
School_Type (Boys)	0.683	0.258	7.015	1	0.008	1.979	1.194	3.281
Location (Urban)	0.343	0.195	3.102	1	0.078	1.409	0.962	2.064
Constant	1.079	0.157	47.089	1	0.000	2.942		

Table 5: Logistic Regression Analysis of Higher Secondary Examination: Overall Results

Discussion

The Wald statistic and the corresponding significance level test, the significance of each of the covariate and dummy independent variables in the model are shown in the above table. If the Wald statistic is significant (i.e., less than 0.05) then the parameter is significant in the model. Of the independent variables, location of school is insignificant, whereas gender of students, medium of instructions and type of schools have significantly affected the results of students in higher secondary examination.

The performances of female students are approximately 3 times higher than that of the performance of male students. As far as medium of instruction is concerned, it can be seen that, the performance of English medium schools are 2.244 times better than that of the Bengali medium schools.

As shown in the above table, type of schools (i.e., girls' and boys' schools) as a whole is a significant factor with p value 0.015 corresponding to examination scores. Also it is observed that, performance of girls' and boys' schools are 0.669 and 1.979 times better than the co-educational schools respectively. Similarly, the performances of urban students are 1.409 times better than the rural students.

Looking at the length of confidence interval of estimated odds, we find that school type (1) i.e., Girls' schools are estimated with 95% confidence having shortest interval length.

Now we consider Matriculation result as a dependent variable, where as gender of students, medium of instruction, school type and location of schools are independent variables.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender (Female)	1.150	0.503	5.222	1	0.022	3.157	1.178	8.463
Medium (English)	-0.125	0.312	0.159	1	0.690	0.883	0.479	1.628
School Type			1.693	2	0.429			
School_Type (Girls)	-0.744	0.575	1.675	1	0.196	0.475	0.154	1.466
School_Type (Boys)	-0.056	0.346	0.026	1	0.872	0.946	0.480	1.863
Location (Urban)	-0.313	0.287	1.187	1	0.276	0.731	0.416	1.284
Constant	2.846	0.273	108.37	1	0.000	17.221		

Table 6: Logistic Regression Analysis of Matriculation Examination: Overall Results

Discussion

Looking at the above Table 6, it is observed that, gender of students' contributes significantly to the model. Where as medium of instruction along with type and location of schools are found to be insignificant in explaining the satisfactory results in matriculation.

Now in particular, the performances of female students are 3.157 times higher than that of the performance of male students. On the other hand for medium of instruction, it can be seen that, the performance of English medium schools are 0.883 times better than that of the Bengali medium schools.

Also from the above table, type of schools (i.e., girls' and boys' schools) as a whole is insignificant factor with p value 0.429 corresponding to examination scores. Hence, it is observed that, performance of girls' and boys' schools are 0.475 and 0.946 times better than the co-educational schools respectively. Similarly, the performances of urban students are 0.731 times better than the rural students.

Looking at the length of confidence interval of estimated odds, we find that location of school is estimated with 95% confidence having shortest interval length.

Comparison of odds ratios of total marks obtained in higher secondary and matriculation examination for the respective set of independent variables:

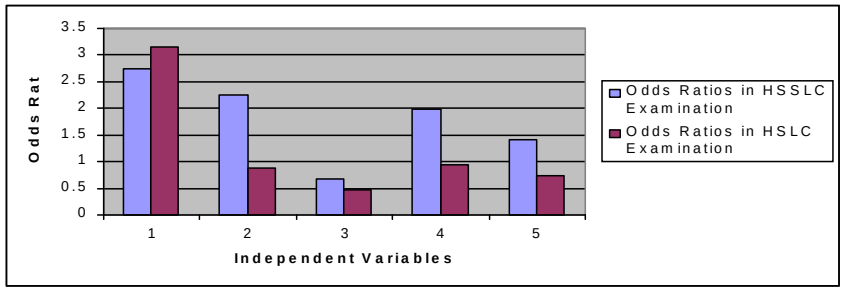

Figure 1: Comparison of Odds Ratios obtained in Higher Secondary (HSSLC) and Matriculation (HSLC) Examination

Discussion

It is apparent from the Figure 1, though female students' are better performer in matriculation, the performance of girls' schools is worse in same examination compared to their performance in higher secondary examination. Similarly English medium and urban students' show the better performance in higher secondary level compared to matriculation. Boys' schools also show better performance in higher secondary stage.

6.2 Analysis (Stage – II)

Here we consider Higher Secondary result as a dependent variable, where as gender of students, medium of instruction, board of examination and location of schools are independent variables.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender (Female)	0.609	0.205	8.839	1	0.003	1.839	1.231	2.747
Medium (English)	-0.635	0.395	2.590	1	0.108	0.530	0.244	1.148
Board			16.696	2	0.000			
Board (TBSE)	-1.351	0.477	8.005	1	0.005	0.259	0.102	0.660
Board (CBSE)	0.994	0.556	3.203	1	0.073	2.703	0.910	8.032
Location (Urban)	0.450	0.195	5.347	1	0.021	1.568	1.071	2.297
Constant	2.582	0.490	27.751	1	0.000	13.225		

Table 7: Logistic Regression Analysis of Higher Secondary Examination: Overall Results

Of the independent variables, gender of students' and location of schools contribute significantly to the model. Board of examination turned out to be highly significant along with TBSE and CBSE whereas CBSE is marginally significant to the model. Only medium of instruction (English medium) shows insignificant result.

The performances of female students are 1.839 times higher than that of the performance of male students. Whereas the performance of English medium schools are 0.530 times better than that of the Bengali medium schools.

As far as board of examination is concerned, it is observed that, TBSE and CBSE boards are 0.259 and 2.703 times better than that of ICSE board. On the other hand, urban students are 1.568 times better than that of rural students.

Looking at the length of confidence interval of estimated odds, we find that, board (1) i.e., Tripura Board of Secondary Education (TBSE) is estimated with 95% confidence having shortest interval length.

Now we consider Matriculation result as a dependent variable, where as gender of students, medium of instruction, board of examination and location of schools are independent variables.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender (Female)	0.755	0.321	5.526	1	0.019	2.128	1.134	3.993
Medium (English)	-0.618	0.579	1.139	1	0.286	0.539	0.173	1.678
Board			1.290	2	0.525			
Board (TBSE)	-0.733	0.672	1.192	1	0.275	0.480	0.129	1.792
Board (CBSE)	-0.103	0.558	0.034	1	0.853	0.902	0.302	2.691
Location (Urban)	-0.288	0.286	1.011	1	0.315	0.750	0.428	1.314
Constant	3.513	0.701	25.120	1	0.000	33.555		

Table 8: Logistic Regression Analysis of Matriculation Examination: Overall Results

From the above Table 8, it has been observed that, only gender contributes significantly to the model. The remaining variables are found to be insignificant.

The performances of female students are 2.128 times higher than that of the performance of male students. On the other hand, performance of English medium schools is 0.539 times better than that of the Bengali medium schools.

As far as board of examination is concerned, it is observed that, TBSE and CBSE boards are 0.480 and 0.902 times better than that of ICSE board. Similarly,

performances of urban students are better than the rural students with 0.750 times.

Looking at the length of confidence interval of estimated odds, we find that, location of school is estimated with 95% confidence having shortest interval length. Comparison of Odds Ratio obtained in higher secondary and matriculation examination for the respective set of independent variables.

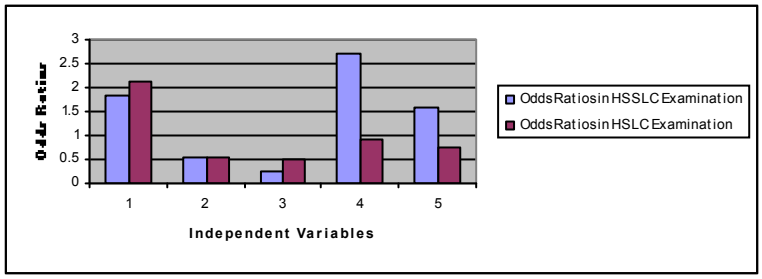

Figure 2: Comparison of Odds Ratios obtained in Higher Secondary (HSSLC) and Matriculation (HSLC) Examination

Discussion

Observing the Figure 2, we may conclude that, female performance is better in matriculation examination than that of higher secondary examination. It is interesting to know that, performance of English medium schools in both the examinations remain same. Similarly performance of TBSE shows that better results in matriculation comparing to higher secondary **e**xamination. Where as, CBSE schools are better performer in higher secondary than that of matriculation examination. In fact urban students are also showing better result in higher secondary examination.

6.3 Analysis (Stage – III)

Here we consider Higher Secondary result as a dependent variable, where as gender of students, medium of instruction, category of school and location of schools are independent variables.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender (Female)	0.564	0.203	7.732	1	0.005	1.758	1.181	2.617
Medium (English)	0.730	0.241	9.192	1	0.002	2.075	1.294	3.326
Category of Schools (Govt)	-0.349	0.190	3.375	1	0.066	0.705	0.486	1.024
Location (Urban)	0.391	0.193	4.099	1	0.043	1.479	1.013	2.159
Constant	1.464	0.178	67.934	1	0.000	4.323		

Table 9: Logistic Regression Analysis of Higher Secondary Examination: Overall Results

From the above Table 9, it is observed that, all the variables except category of school are found to be significant.

The performances of female students are 1.758 times higher than that of the performances of male students. As far as medium of instruction is concerned, 2.075 times better performance of English medium school is observed, than that of the Bengali medium schools.

On the other hand, the performances of Govt. schools comparing to the others i.e., non-Govt. schools are better (with odds ratios 0.705). Similarly, the performances of urban students are 1.479 times better than the rural students.

Looking at the length of confidence interval of estimated odds, we find that category of school is estimated with 95% confidence having shortest interval length. Here we consider Matriculation result as a dependent variable, where as gender of students, medium of instruction, category of school and location of schools are independent variables.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender (Female)	0.748	0.320	5.467	1	0.019	2.113	1.129	3.955
Medium (English)	-0.117	0.310	0.143	1	0.706	0.890	0.485	1.632
Category of School (Govt)	-0.438	0.288	2.322	1	0.128	0.645	0.367	1.134
Location (Urban)	-0.329	0.287	1.319	1	0.251	0.719	0.410	1.262
Constant	3.042	0.295	106.53	1	0.000	20.950		

Table 10: Logistic Regression Analysis of Matriculation Examination: Overall Results

In third stage analysis, gender is also the only variable found to be significantly contributing to the model.

The performances of female students are 2.113 times higher than that of the performances of male students. Whereas performance of English medium school is 0.890 times better than that of the Bengali medium schools.

On the other hand, it is observed that, the performances of Govt. schools are 0.645 times better than, the non-Govt. schools. Similarly, the performances of urban students are 0.719 times better than the rural students.

Looking at the length of confidence interval of estimated odds, here we also find that category of school is estimated with 95% confidence having shortest interval length.

Comparison of Odds Ratios obtained in higher secondary and matriculation examination for the respective set of independent variables.

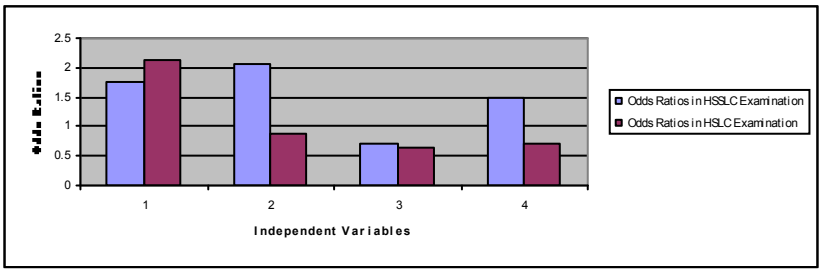

Figure 3: Comparison of Odds Ratios obtained in Higher Secondary (HSSLC) and Matriculation (HSLC) Examination

Discussion

Observing the Figure 3, it may be point out that, females show better performance in matriculation compared to higher secondary examination. But performance of English medium schools turned out to be better in higher secondary examination. On the other hand, Govt. schools are showing almost same performance in both the examination. Urban schools shows 2 times better performance in higher secondary examination compared to matriculation.

Overall Conclusion

From the above observations in all the three stages, we may conclude that, females are always showing best performances in both the examinations. But their performance is found to be still better in matriculation as shown in our analysis in all three stages. The performances of English medium schools are found to be satisfactory in higher secondary examination. Similarly urban schools always show better performance in higher secondary examination.

Acknowledgement

The author acknowledges the University Grants Commission (UGC), New Delhi, India for providing financial support through a Minor Research Project (MRP). The author also expresses his gratitude to the editor and the reviewers of Journal of Reliability and Statistical Studies (JRSS) to improve the paper through constructive suggestions.

References

1. Agresti, A. (1996). An Introduction to Categorical Data Analysis, John Wiley and Sons, Inc.
2. Hosmer, D. and Stanley, L. (1989). Applied Logistic Regression, John Wiley and Sons, Inc.
3. Menard, S. (1995). Applied logistic regression analysis (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–106), Thousand Oaks, CA: Sage.

4. Menard, S. (1995). Applied Logistic Regression Analysis. Sage Publications.Series: Quantitative Applications in the Social Sciences, No. 106.
5. Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis, The American Statistician, 54(1), p. 17–24.
6. Saha, G. and Sarmah, P. (2010). Statistical Analysis of School Examination Result with Special Reference to the State of Tripura: North-East India, Journal of Statistics Sciences, 2, p. 111-121.
7. Saha, G. and Sarmah, P. (2012): Stochastic Modeling of the Grading Pattern in Presence of the Environmental Parameter, Electronic Journal of Applied Statistical Analysis (Accepted).
8. Sarma, R. and Sarmah, P. (1999). A Stochastic Modeling on Grading System, In Proceedings of the Second International Conference on Operations and Quantitative Management in the Global Business Environment (ICOQM), Ahmedabad, India, 3–6th January 1999, p. 276–281.
9. Sarma R. and Sarmah P. (1999). Analysis of Results Based on Grades, In Proceedings of the Second International Conference on Operations and Quantitative Management in the Global Business Environment (ICOQM), Ahmedabad, India, 3–6th January 1999, p. 282–290.
10. Tabachnick, B. and Linda, F. (1996). Using Multivariate Statistics, Third edition. Harper Collins.