Volume-4, Issue-7, July-2015 • ISSN No 2277 - 8160

Research Paper Statistics

Determination of Students' Pass Rate of a Particular Examination: A Regression Modeling Scenario (GLM) Approach

Jugal Gogoi Research Scholar,Department of Mathematics,Dibrugarh University,Dibrugarh-786004, Assam(India).

Navajyoti Tamuli Research Scholar, , Nemuguri, Sivasagar-785674. Assam(India).

ABSTRACT

The main purpose of this study was to find out the isolated and combined effect of brisk walking and aerobic exercise on muscular strength and endurance and cardiovascular endurance among middle aged men For this study sixty middle aged men selected from Gudiathamvellor District. The age of the subjects ranged from 35 to 45 years. The selected Sixty subjects were divided into four equal groups(n-15), Experimental Group A named as Briskwalking group, experimental groups B named as Aerobic exercise group, experimental group C named combined (brisk walking and aerobic exercise) group and Group D acted as control group. Pre-test was conducted on muscular strength and endurance and Cardio vascular Endurance for all groups. The measurements were carefully regarded in their respective unit as pre-test score. After pre-test Experimental Group A was treated with briskwalking, Experimental Group B was treated with Aerobic exercise and Experimental group C treated with combined group exercise for five days per a week for the period of twelve weeks. Whereas Group D not treated with any specific training,they kept as control group. After twelve weeks of training post-test was conducted on muscular strength and endurance and cardiovascular endurance measured and carefully recorded as post-test score. The collected pre and post-test data from the three groups were statistically analyzed by analysis of covariance. Whenever the F ratio was found to be significant for adjusted post-test mean, Scheffe's test was followed as a post hoc test to determine the level of significant difference between the paired means. The level of significance was fixed at 0.05 level of confidence.

KEYWORDS : GLM, Linear Model, regression, Binomial Distribution, Normal Distribution, Response Variable

Introduction:

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data. The term *regression* literally means "*step back towards the average*" [2]. In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called

dependent variable or regressed or explained variable and the variable which influences the values or is used for prediction is called independent variable or regressor or predictor. When the response variable depend only one regressor, and the equation of the model is linear then this type of regression is known as *simple linear regression*. A regression model that involves more than one regressor variable is called a *multiple regression model*. Generally if response related to regressor or predictor variables, the model

is called a *multiple linear regression model* with regressors since the equation of this model is linear. The parameters , are called regression coefficients. This model describes a hyper plane in the - dimensional space of the regressor variables. The parameter represents the expected change in the response per unit change in when all of the remaining regressor variables are held constant. For this reason the parameters are often called partial regression coefficients. The method of least squares can be used to estimate the regression coefficients. The linear regression models provide a rich and flexible framework that suits the needs of many analysts. However, linear regression models are not appropriate for all situations. There are many problems in real world where the response variable and the predictor variables are related through a known non-linear function. This leads to a non-linear regression model. Any regression model that is not linear is a non-linear regression model. When the method of least squares is applied to such models, the resulting normal equations are non-linear and in general, difficult to solve. The generalized linear model (GLM) is a unification of both linear and non-linear regression models that also allows the incorporation of non-normal response distributions. In a GLM, response variable distribution must only be a member of the exponential family[3][4], which includes the normal, Poisson, binomial, exponential and gama distributions as members. Furthermore, the normal-error linear model is just a special case of the GLM, so in many ways, the GLM can be thought of as a unifying approach to many aspect of empirical modeling and data analysis. GLMs relate a variable (called response variable) which we want to predict, to variables or factors (called predictors, covariates or independent variables) about which we have information. In order to do this, it is necessary first to define the distribution of the response. Then the covariates can be related to the response allowing for the random variation of the data. In the present study we are setting up a model to predict the pass rate of a particular student in a particular examination using GLM framework, considering the response follows Binomial distribution. It determines many factors that affect whether a student is likely to pass or not. 110 class-x students' information of four schools under Jatiya Bidyalaya Sanstha of Sivasagar district is used for demonstration.

2. Composition of GLM .

Generalized Linear Model (GLM) consists of three components:

1. A random component, specifying the conditional distribution of the response variable, (for the i^{th} of n independently sampled observations), given the values of the explanatory variables in the model. Although Nelder and Wedderburn[5], originally formulated that, the distribution of must be a member of the exponential family, later works on GLIM has extended to multivariate exponential families such as multinomial distribution and to certain non-exponential distributions (e.g. two parameter negative-binomial distribution) and in certain cases the distribution of response is not specified completely.

2. A linear predictor - that is a linear function of regressors

As in the linear model, the regressors are prespecified functions of the explanatory variables which include quantitative explanatory variables, transformations of quantitative explanatory variables, polynomial regressors, dummy regressors, interactions and so on. One of the advantages of GLIM is that the structure of the linear predictor is the familiar structure of a linear model.

3. A smooth and invertible linearizing link function, which transforms the expectation of the response variable,, to the linear predictor,

Because of the link function is invertible. Thus GLM is a linear model for a transformation of the expected response or a non linear regression model for the response. The inverse link is called the mean function. Although the selecting of a link function makes the regression of Y on Xs is linear, a promising link will remove restrictions on the range of the expected response.

3. Binary Response Variable and Link Functions:

For binomial data, represents the observed proportion of 'success-

es' in independent binary trials; thus can take any of the values. If the response variable in a regression problem takes only two possible values, 0 and 1 (mostly observing from qualitative response) then the response variable is a Bernoulli random variable, indicating the shape of the response function is non-linear. Binomial data also encompass binary data, where all the observations represent trial and consequently is either 0 or 1. The expectation of the response is then the probability of success. As a probability, is confined to the unit interval [0, 1]. The logit and probit links map this interval to the entire real line, from. Similarly if the response Y is a count, taking on only non-negative integer values, 0,1,2, …, and consequently is an expected count, which (though not necessarily an integer) is also non-negative, the log link maps to the whole real line. So the selection of link function does not entirely determined by the range of the response variable.

In the present study the students' performance (response; pass = 1 and fail = 0) is considered to follow Binomial distribution. The commonly employed link functions for Binomial response are Logit, Probit, Log-Log and Complementary Log-Log etc, of which Logit link function is considered in the present study for regression diagnostics in GLM framework. The log-log link can be obtained from the complementary log-log link by exchanging the definitions of success and failure. The link functions and their inverses for Binomial responses are given below in **Table. (1)**. Here, is the expected value of the response; η_i is the linear predictor; and is the cumulative distribution functions of the standard normal distribution

Table.(1): Some Common Link Functions & Their Inverses for Binomial Response

Link	
Logit	
Probit	
Log-Log	]
Complementary Log-Log	]

4. Estimating and Testing GLM

In GLM, data are fitted by the method of maximum likelihood, which provides estimates of regression coefficients as well as estimated standard errors of the coefficients. For testing the null hypothesis , Wald Chi square statistic is computed, where is the standard error of the estimated coefficient. Under the null hypothesis, follows a standard normal distribution. Hypothesis testing in GLM is based on likelihood ratio tests. The likelihood ratio approach leads to a statistic named as deviance. The deviance of a model compares the log-likelihood of the fitted model of interest to the log-likelihood of the saturated model, i.e. model has exactly n parameters that fits the sample data perfectly. The value of the log likelihood function for the fitted model can never exceed the value of the log-likelihood function for the saturated model, as the fitted model contains fewer parameters. The deviances for a GLM is defined as,)

Here is the maximized likelihood under the model to be fitted and is the maximized likelihood under the saturated model, which dedicates one parameter to each observation and consequently fits the data as closely as possible. This residual deviance of GLIM is the generalization of residual sum of squares for a linear model. Large value of the model deviance indicates that the model is not correct, while a small value implies that the fitted model fits the data. The test criteria is,

if, ,conclude that the fitted model is adequate,
if, ,conclude that the fitted model is not adequate,

The expression for binomial deviance can be figured out as follows. Let denote the m.l.e. of under the model of interest, and let denote the m.l.e. under the saturated model. Then the deviance,

All the terms involving cancel out. Collecting terms on and , it is found that,

This deviance is used to test hypothesis on subsets of the model parameters. The regression model in GLIM context can be written as,

Here the full model has parameters, contains of these parameters, contains of these parameters, and the columns of the matrices and contain the variables associated with these parameters. To test a reduced model, the test hypothesis becomes,

And then the reduced model becomes,

To test the goodness fit of the reduced model, deviance for reduced model is compared with the deviance full model, as the deviance for the reduced model is larger than the deviance of full model, due to fewer parameters of reduced model than full model. Yet, if is not much larger than the full model deviance, then it indicates that the reduced model is a good fit as the full model, so it is likely that the parameters in are equal to zero. The difference in deviance is - , which has degrees of freedom. If the null hypothesis is true and is large then difference in deviance - has a Chi-square distribution with degrees of freedom. Difference in deviance is also known a partial deviance. Then the test statistic and decision criteria are,

If, - , reject the null hypothesis and if,
- , accept the null hypothesis.

5. Illustration: Students' performance Data

For illustration seven variables are considered as predictors for estimating students' pass rate. They are Father's education, Mother's education, Quarter examination mark, Number of assignment submitted before the examination, either he attend tutorial or not, Family size and family type. Altogether110 primary data on above mentioned variables of students' pass/fail data are collected from class-x students' information of four schools under Jatiya Bidyalaya Sanstha of sivasagar district Assam. The dichotomous outcome variable of interest in this study was "pass" or "fail" on a particular examination. Passing the examination required that a student achieve a total 60% and above as the minimum passing standard. For analysis, this variable was coded as 0 = failure and 1 = pass. Here the students' result of a particular examination (response; pass=1 and fail=0) is considered to follow Binomial distribution. Regression diagnostics of students' performance data in GLM is considered due to distribution of response is non linear and many predictors are categorical. The commonly employed link functions for Binomial response are Logit, Probit, Complementary Log-Log and Log-Log, of which logit link is employed here for regression diagnostics of students' performance data. In GLM, hypothesis testing is based on likelihood ratio tests as well as in deviance statistics. Results are summarized in the **Table.(2)** and **Table.(3)** in GLM framework.

The deviance statistics for all possible subset models, full models including intercept models for logit link function are displayed in **Table.(2)**, assuming response data follows Binomial distribution. In generalized linear model, model deviance is used to test hypotheses on subsets of the model parameters. To investigate the potential usefulness of various subsets models, all possible 544 models considering logit link functions fitted in GLM regression setting. The deviance statistics with respective degrees of freedom of each models show that reduced models have higher deviances than the full models as expected. To select one best model among all possible models, we divide the model deviance by its degrees of freedom . If the ratio is close to unity, then the model is considered adequate. Although we expect nine factor for analyzing students' pass rate, but using the ratio of model deviance and respective degrees of freedom we get a seven variable model which is considered adequate. Now, to test the hypothesis on subsets of model parameters, a seven variable model (FE, ME, NA, QM, T, FS, FT) with logit link and deviance 69.503 is considered as a model to be tested for lowest difference in deviance, i.e., which is less than 10.60. So for students' pass rate model, the value of the test statistic becomes,

- = 10.33 (i.e. 10.60),

which indicates that the null hypothesis of some parameters of the saturated model are zero- is accepted. Thus on the basis of the test statistic and decision criteria, students' performance data- seven variable reduced model (FE, ME, NA, QM, T, FS, FT) is considered as the best fitted model. Furthermore, the logit link function is chosen as the promising link for binomial response variable, where predictors are categorical as well as quantitative. **Table. (3)** is reported all the information regarding parameter estimates of best fitted full model for

students' performance data of the present piece of work. **Table.(3)** also displays the estimated parameters, their standard errors, value of test statistic including confidence intervals for seven variable best selected reduced model of the present study.

Table.(2): Likelihood Ratio Statistics: Students' Performance Data
Response Variable: Examination Result
Probability Distribution: Binomial

Sl. No.	Model	Deviance()	d.f.	/d.f.
1	FULL	59.173	97	0.61003
2	INTERSEPT	34.167	108	0.31636
3	ME	25.571	107	0.23898
4	FE	25.700	107	0.24019
5	MO	35.935	106	0.33901
6	FO	33.709	106	0.31801
7	FS	35.141	107	0.32842
8	FT	32.736	107	0.30594
9	QM	71.009	107	0.66364
10	T	34.855	107	0.32575
11	NA	22.207	107	0.20754
12	ME, FE	23.398	106	0.22074
13	ME, MO	26.426	105	0.25168
14	ME, FO	24.833	105	0.23651
15	ME, FT	24.547	106	0.23158
16	ME, T	26.535	106	0.25033
17	ME, NA	18.203	106	0.17173
18	ME, QM	70.316	106	0.66336
19	ME, FS	26.533	106	0.25031
20	FE, MO	27.166	105	0.25872
21	FE, FO	25.734	105	0.24509
22	FE, FT	23.103	106	0.21795
23	FE, T	26.495	106	0.24995
24	FE, NA	16.027	106	0.1512
25	FE, QM	70.591	106	0.66595
26	FE, FS	26.549	106	0.25046
27	MO, FO	35.242	104	0.33887
28	MO, FT	34.486	105	0.32844
29	MO, T	36.680	105	0.34933
30	MO, NA	22.936	105	0.21844
31	NA, QM, T	70.536	105	0.67177
32	ME, NA, QM	70.007	105	0.66673
33	ME, NA, QM, T	69.878	104	0.6719
34	FE, NA, QM, T	70.021	104	0.67329
35	FE, ME, QM, T, FS	69.661	103	0.67633
36	FE, ME, NA, QM, T, FS	69.513	102	0.6815
37	FE, ME, NA, QM, T, FS, FT	69.503	101	0.68815

Model Coefficients of Students' Pass Rate Model

Table: (3): Reduced seven Variable Best Fitted Model / Link: Logit /Response: Examination Result

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	15.436	4.2870	7.034	23.838	12.965	1	.000	5055801.993	1134.307	2.253E10
[Mothers_ Education=.00]	.367	.7853	-1.173	1.906	.218	1	.641	1.443	.310	6.725
[Mothers_ Education=1.00]	0[a]	.	.	.	.	.	.	1	.	.
[Fathers_ Education=.00]	.281	.7345	-1.158	1.721	.147	1	.702	1.325	.314	5.589
[Fathers_ Education=1.00]	0[a]	.	.	.	.	.	.	1	.	.
[Tutorial=.00]	.304	.7733	-1.212	1.819	.154	1	.695	1.355	.298	6.167
[Tutorial=1.00]	0[a]	.	.	.	.	.	.	1	.	.
[Family_Type=.00]	.104	1.0458	-1.946	2.154	.010	1	.921	1.110	.143	8.619
[Family_Type=1.00]	0[a]	.	.	.	.	.	.	1	.	.
Number_of_ Assignment	-.153	.4091	-.955	.648	.140	1	.708	.858	.385	1.913
Quarter_Marks	-.238	.0518	-.339	-.136	21.038	1	.000	.788	.712	.873
Family_Size	-.065	.1882	-.434	.304	.120	1	.729	.937	.648	1.355
(Scale)	1[b]									
a. Set to zero because this parameter is redundant.										
b. Fixed at the displayed value.										

The seven predictors of the best selected reduced model for which estimated parameters exist father's education, mother's education, attendance of tutorial, family type, number of assignment submitted, half year examination mark and size of family. So on the basis of the results given in the **Table.(3)**, our Binomial regression model with Logit link function for Students' performance data will be as follows,

Where,

1, if father education is up to Higher Secondary = 0, if father education is above Higher Secondary, if mother education is up to Higher Secondary 0, if mother education is above Higher Secondary is the half year examination mark is the number of assignment by the student 1, if a student attend tutorial 0, otherwise is the size of household if a student belongs to single family. 0, if he belongs to joint family The deviance for this model is 69.503 and with 101 d.f. So our conclusion is that the model is an adequate fit to the data. Besides, the logit link is considered as a suitable link for binomial responses with quantitative as well as categorical predictors.

6. Conclusion

A family of regression models called generalized linear model is actually a unifying approach to regression and experimental design models, uniting the usual normal-theory regression models and non linear regression models. Formulated by Nelder and Wedderburn (1972), GLM has the assumption that the distribution of response must be a member of exponential family of probability distributions. In addition to regression models for continuous response variables, models for rate and proportions, binary, ordinal and multinomial variables and counts can be handled as GLMs. Its approach to general theoretical framework for many commonly encountered statistical models is attractive.

Student performance on a particular exam is a function of several inputs. Scores are affected by the students' academic aptitude, the socio-economic characteristics of the students' households, and the characteristics of the students' schools. There are so many social, demographic and economic factors which are influence students performance. To study these influences quantitatively and analytically GLM regression performs best as many socio-economic and demographic data are continuous as well as categorical. In our students' performance data father's education, mother's education, attendance of tutorial, family type are categorical and number of assignment submitted, half year examination mark and size of family are quantitative variables. Considering response follows binomial distribution, regression diagnostics are carried out in GLM settings. Comparing number of subset binomial regression models with logit link function, one seven variable model is selected as adequate fit to the data.

REFERENCES

[1] Montgomery, D.C., Peck, A.E., Vining, G.G.(2001): Introduction to Linear Regression Analysis; John Wiley & Sons. [2] S.C. Gupta & V. K. Kapoor: Fundamental of Mathematical Statistics; Sultan Chand &sons publication. [3] McCullah,P. , Nelder, J.A.(1983): Generalized Linear Models, London; Chapman & Hill. [4] Singh, S., (1995): Self Help Group in Indian Agribusiness- Replications from Case Studies; Artha Vijnana 37(4): 380-388. [5] Nelder, J.A., Wedderburn, R.W.M. (1972): Generalized Linear Interactive Models; Journal of the Royal Statistical Society, Ser.A135: pp.370-384.