Discriminating Between Response Scores In A Diagnostic Test: A Dummy Variable Regression Approach

Oyeka I.C.A, Awopeju K.A, Efobi C.C., Onyiaorah A.A.

Abstract: This paper proposes a statistical method for the analysis of diagnostic screening test results reported as quantitative scores, using dummy variable multiple regression techniques. The proposed method develops estimates of the expected test scores for subjects whose tests scores are critically below the minimum normal score; those subjects whose test scores are critically above the maximum normal score; those subjects whose test scores are either marginally below the minimum normal score or marginally above the maximum normal score; and for those subjects whose test results or scores are normal. Test statistics are also developed for testing the existence of any significant difference between the expected scores by these various groups of subjects. The proposed method which may enable health practitioners in statistically discriminating screened subjects, for specific health management programs, by identified risk groups relative to the normal group, is illustrated with some data.

Keyword: Dummy Variable, Multiple Regression, Classification, Discrimination, Probability Density Function, Parameter, Estimate.

1. Introduction

There are often many possible outcomes in a diagnostic test. Results from a diagnostic screening test may indicate that a randomly selected subject tests negative, that is, does not possess the condition of interest, meaning that the condition being investigated is definitely absent. The result may also indicate that the subject has tested mildly positive that is the test is either not definitely positive or positive but not critical. Finally the result may indicate that the test is definitely positive or positive and critical showing that the condition is definitely present. There are mainly four broad groups here that may be of public health interest. These include those subjects whose scores in the diagnostic test are critically below the minimum normal level or critically above the maximum normal level; those subjects whose scores are within the normal levels and those whose scores are either marginally below or marginally above the normal levels. The later groups of subjects may possibly be pooled together for some common health management such as medical or some other counseling. Research interest may be in determining whether the proportions of the population responding under this various categories are statistically different. In this paper we propose to use the dummy variable regression method in which membership in the various categories are coded 1s and 0s to handle this problem. It is here assumed that the responses are quantitative, that is, are numerical scores measured on the real line.

2. The Proposed method

Suppose a random sample of 'n' subjects is drawn from a certain population for screening in a diagnostic test for some condition of interest. Let y_i be the score of the i^{th} subject in the diagnostic test, for i = 1, 2, …, n. We assume that values of y_i in the interval (c_2, c_3) indicate that the i^{th} subject has definitely, a negative response, that is, the subject has a normal score or the condition of interest is absent. Values in the intervals (c_1, c_2) or (c_3, c_4) indicate that the i^{th} subjects score is either marginally below the lowest normal score or marginally above the highest normal score. Values of y_i that are less than c_1 or greater than c_4 indicate that the i^{th} subject score is either critically below the minimum normal score (c_2) or critically above the maximum normal score (c_3); where c_1, c_2, c_3 and c_4 are non negative real numbers with $c_1 < c_2 < c_3 < c_4$. Thus there are four categories in which subjects may be grouped according to their scores in the diagnostic test. These are those subjects whose scores are critically below the minimum normal score, that is those with scores $y_i < c_1$; those whose scores are critically above the maximum normal score, that is those with scores $y_i > c_4$; those with normal scores that is those with scores y_i such that $c_2 < y_i < c_3$; and those whose scores are either marginally below the minimum normal score that is those with scores that are marginally above the maximum normal score, that is those with scores y_i such that $c_1 \le y_i \le c_2$, and those whose scores are marginally above the maximum normal score, that is those with scores y_i such that $c_3 \le y_i \le c_4$, for i = 1, 2, …, n. Now consistent with the use of dummy variable representation in regression models in which each parent variable is represented by one dummy variable less than the number of its categories, (Boyle 1970; Neter and Wasserman 1983; Oyeka 1993) we here use three dummy variables of 1's and 0's to represent the four categories in to which subjects are grouped by y_i, their scores in the diagnostic test;
Specifically let

$$x_{i1} = \begin{cases} 1, \text{if } y_i < c_i \\ 0, \text{Otherwise} \end{cases}$$

$$x_{i2} = \begin{cases} 1, \text{if either } c_1 \le y_i \le c_2 \text{ or } c_3 \le y_i \le c_4 \\ 0, \text{Otherwise} \end{cases}$$

$$x_{i3} = \begin{cases} 1, \text{if } c_2 < y_i < c_3 \\ 0, \text{Otherwise} \end{cases} \qquad \text{.....} \qquad 1$$

for = i =1, 2, …, n.

A linear regression model of the scores y_i on x_{i1}, x_{i2}, and x_{i3} may be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \quad \ldots \quad 2$$

for i =1, 2, …, n.

where β_0, β_1, β_2, and β_3, are regression coefficients and e_i are error terms uncorrelated with the $x_{ij's}$ An alternative expression of Eqn 2 in matrix form is

$$\underline{y} = X\underline{\beta} + \underline{e} \quad \ldots\ldots\ldots \quad 3$$

where $\underline{y}$ is an n x 1 column vector of scores, X is an n x p design matrix of 1s and 0s, $\underline{\beta}$ is a p x 1 column vector of regression coefficients and $\underline{e}$ is an n x 1 column vector of error terms uncorrelated with X, where p = 4 the number of parameters (regression coefficients) in the model. The method of least squares may be used with Eqn 2 or 3 to obtain unbiased estimates of the regression parameters as

$$\underline{\hat{\beta}} = \underline{b} = (X'X)^{-1}X'Y \quad \ldots\ldots\ldots \quad 4$$

where $(X'X)^{-1}$ is the matrix inverse of $X'X$.

The resulting estimated regression model is

$$\underline{\hat{y}} = X\underline{b} \quad \ldots\ldots\ldots \quad 5$$

To check whether the regression model fits we test the null hypothesis

$H_0 : \underline{\beta} = \underline{0}$ or $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus H_i : not all the $\beta_{js} = 0$ … 6

for some j =1, 2, 3.

The null hypothesis is tested using the following ANOVA table (table 1) based on the F-distribution.

Table 1: ANOVA table for Equation 3

Source of variation	Sum of Square SS	Degrees of Freedom DF	Mean Sum of Squares MS	F-Ratio
Regression	$SSR = \underline{b}'X'Y - n\bar{Y}^2$	$p - 1 = 4 - 1 = 3$	$MSR = \frac{SSR}{3}$	$F = \frac{MSR}{MSE}$
Error	$SSE = Y'Y - \underline{b}'X'Y$	n – p = n – 4	$MSE = \frac{SSE}{n-4}$	
Total	$SST = Y'Y - n\bar{Y}^2$	n – 1		

H_0 is rejected at the α level of significance if

$$F \geq F_{1-\alpha;3,n-4} \quad \ldots\ldots\ldots \quad 7$$

Otherwise H_0 is accepted where $F_{1-\alpha;3,n-4}$ is the critical value of the F distribution with 3 and n – 4 degrees of freedom for a specified α – level. If the model fits, that is if H_0 is rejected in which case not all β_js = 0 then we can use these β_js to discriminate between response score of subjects. Now the estimated response score when the ith subject test result is critically below the minimum normal level or score is obtained by setting $x_{i1} = 1$ and $x_{i2} = x_{i3} = 0$, in equation 2 as $\beta_0 + \beta_1$ which is estimated as

$$\hat{\beta}_0 + \hat{\beta}_1 = b_0 + b_1 \quad \ldots\ldots \quad 8$$

Similarly the expected response score if the i^{th} subject's test result is either marginally below the minimum normal level or marginally above the maximum normal level or is normal are respectively

$\beta_0 + \beta_2$ when $(x_{i1} = 0, x_{i2} = 1, x_{i3} = 0)$ and $\beta_0 + \beta_3$ when $(x_{i1} = x_{i2} = 0; x_{i3} = 1)$ which are estimated respectively by

$$\hat{\beta}_0 + \hat{\beta}_2 = b_0 + b_2 \text{ and } \hat{\beta}_0 + \hat{\beta}_3 = b_0 + b_3 \ldots 9$$

The expected (mean) test score for subjects whose test scores are critically above the maximum normal score ($x_{i1} = x_{i2} = x_{i3} = 0$ in Eqn 2) β_0 estimated as

$$\hat{\beta}_0 = b_0 \quad \ldots\ldots \quad 10$$

Interest may be in testing the null hypothesis of no differences between these expected test scores responses having rejected the null hypothesis that they are all equal. Thus the expected differences in the mean scores of subjects whose test results are critically below the minimum normal score and those whose scores are marginally below the minimum normal score or marginally above the maximum normal score is

$(\beta_0+\beta_1)-(\beta_0+\beta_2)=\beta_1-\beta_2$ which is estimated by b_1-b_2

Hence the null hypothesis that these two mean responses are equal is equivalent to the null hypothesis

$$H_0:\beta_1-\beta_2=0 \; versus \; H_1:\beta_1-\beta_2\neq 0 \quad ...11$$

which may be tested using the test statistics

$$t=\frac{b_1-b_2}{se(b_1-b_2)}=\frac{b_1-b_2}{\sqrt{(c_{11}+c_{22}-2c_{12})MSE}} \quad 12$$

which has the student t distribution with

n – 4 degrees of freedom, where c_{ij} is the entry in the i^{th} row and j^{th} column of $(X'X)^{-1}$, the matrix inverse of $X'X$ for i = 1, 2, 3 and j = 1, 2, 3.

H_0 is rejected at the α level of significance if

$t \geq t_{1-\alpha/2;n-4}$, otherwise H_0 is accepted ..13

Where $t_{1-\alpha/2;n-4}$ is the critical value of the t distribution with n – 4 degrees of freedom at a specified α level. Similarly the null hypothesis that the mean scores of subjects whose test results are critically below the minimum normal score and those subjects whose test results are normal $H_0:(\beta_1=\beta_3)$ is tested using the test statistic

$t=\frac{b_1-b_2}{se(b_1-b_2)}=\frac{b_1-b_2}{\sqrt{(c_{11}+c_{22}-2c_{12})MSE}}$ which also has a t distribution with n – 4 degrees of freedom The null hypothesis that the mean score of subjects whose test results are critically below the minimum normal score and those whose test results are critically above maximum normal score is equivalent to testing the null hypothesis,

$H_0:\beta_0+\beta_1-\beta_0=0 \; or \; H_0:\beta_1=0$ versus

$$H_1:\beta_1\neq 0 \quad 14$$

This null hypothesis is tested using the test statistic

$$t=\frac{b_1}{se(b_1)}=\frac{b_1}{\sqrt{c_{11}MSE}} \quad \ldots\ldots\ldots \quad 15$$

which also has a t distribution with n – 4 degrees of freedom Similarly testing the null hypothesis that the mean score of subjects who respond normal is the same as the mean score of subjects whose test results are critically above the maximum normal score is equivalent to testing

$$H_0:\beta_3=0 \text{ versus } H_1:\beta_3\neq 0 \quad \ldots\ldots \quad 16$$

Using the test statistic

$$t=\frac{b_3}{se(b_3)}=\frac{b_3}{\sqrt{c_{33}MSE}} \quad \ldots\ldots\ldots \quad 17$$

which also has the t distribution with n – 4 degrees of freedom. Note that if the researcher is also interested in distinguishing between subjects whose test results are marginally below the minimum normal score and subjects whose test results are marginally above the maximum normal score, the two groups may be treated as two separate groups so that we now have five instead of four groups to consider. These five groups are then represented by four dummy variables by including a fourth dummy variable x_{i4} in the specification of Equation 1. The regression model (Eqn 2) is also modified accordingly so that the number of regression parameters in the model is now $p=5$ instead of 4. With these modifications, analyses would then proceed as usual. Results from the proposed method may be compared with the results that would have been obtained had the data on the subjects responding under the various response categories been treated as a one-way analysis of variance problem.

3. Illustrative Example

We illustrate the above procedure using data on the Low Density Lipo-protein (LDL) levels of a random sample of 54 subjects in a certain community (Table 2)

Table 2: Low Density Lipo-Protein (LDL) levels of 54 randomly selected subjects

S/No	LDL levels	S/No	LDL levels	S/No	LDL levels	S/No	LDL levels	S/No	LDL levels	S/No	LDL levels
1	1.97	10	6.88	19	3.95	28	1.24	37	1.31	46	1.55
2	5.14	11	1.43	20	1.75	29	4.41	38	2.07	47	4.74
3	5.7	12	1.56	21	1.89	30	4.20	39	1.08	48	5.59
4	1.2	13	0.87	22	10.2	31	3.62	40	4.37	49	6.76
5	5.3	14	0.55	23	1.68	32	1.25	41	5.54	50	3.96
6	3.0	15	0.99	24	1.59	33	2.02	42	2.88	51	2.21
7	1.6	16	1.11	25	3.92	34	4.53	43	1.03	52	3.50
8	1.14	17	1.34	26	2.75	35	3.08	44	4.34	53	1.09
9	4.1	18	2.51	27	3.15	36	1.30	45	1.27	54	3.66

Normal LDL Range (1.68, 4.53) ≡ (C_2, C_3)

To illustrate the proposed method using the data of Table 2 we note that the normal range of Low Density Lipo-protein (LDL) level is here specified to be (1.68, 4.53) equivalent to our (C_2 and C_3). Hence subjects whose Low Density Lipo-protein level are within this range may be assumed to be normal or to respond negative. We further assume for illustrative purposes but without loss of generally that subjects whose low density lip-protein levels are 0.5 units less than the lower normal level of 1.68 (= C_2) or 0.5 units above the upper normal level of 4.53 (=C_3) are mildly positive; and subjects whose Low Density Lipo-protein levels are more than 0.5 units less than the lower normal level or 0.5 units greater than the upper normal level are here considered as having critical, definitely positive response to the diagnostic test. Therefore c_1 = 1.18, c_2 = 1.68, c_3 = 4.53, c_4 = 5.03. Applying equation 1 to the Low Density Lipo-Protein (LDL) levels in Table 2 assuming the above specified limits, we obtain the following design matrix X for these data (Table 3)

Table 3: Design Matrix X for the Data of Table 2

S/No	LDL Level y_i	x_{i1} $y_i<1.18$	x_{i2} $1.18\leq y_i\leq 1.68$ or $4.53\leq y_i\leq 5.03$	x_{i3} $1.68<y_i<4.53$
1	1.97	0	0	1
2	5.14	0	0	0
3	5.7	0	0	0
4	1.2	0	1	0
5	5.3	0	0	0
6	3.0	0	0	1
7	1.6	0	1	0
8	1.14	1	0	0
9	4.1	0	0	1
10	6.88	0	0	0
11	1.43	0	1	0
12	1.56	0	1	0
13	0.87	1	0	0

14	0.55	1	0	0
15	0.99	1	0	0
16	1.11	1	0	0
17	1.34	0	1	0
18	2.51	0	0	1
19	3.95	0	0	1
20	1.75	0	0	1
21	1.89	0	0	1
22	10.2	0	0	0
23	1.68	0	1	0
24	1.59	0	1	0
25	3.92	0	0	1
26	2.75	0	0	1
27	3.15	0	0	1
28	1.24	0	1	0
29	4.41	0	0	1
30	4.20	0	0	1
31	3.62	0	0	1
32	1.25	0	1	0
33	2.02	0	0	1
34	4.53	0	1	0
35	3.08	0	0	1
36	1.30	0	1	0
37	1.31	0	1	0
38	2.07	0	0	1
39	1.08	1	0	0
40	4.37	0	0	1
41	5.54	0	0	0
42	2.88	0	0	1
43	1.03	1	0	0
44	3.34	0	0	1
45	1.27	0	1	0
46	1.55	0	1	0
47	4.74	0	1	0
48	5.59	0	0	0
49	6.76	0	0	0

50	3.96	0	0	1
51	2.21	0	0	1
52	3.50	0	0	1
53	1.09	1	0	0
54	3.66	0	0	1

Applying Equation 5 to the data of Table 3 we obtain the fitted regression model

$$\hat{y} = 5.90 - 4.92x_{i1} - 4.06x_{i2} - 2.70x_{i3} \quad .. 18$$

and the corresponding analysis of variance table (table 4) is

Table 4: ANOVA table for Equation 3

Source of variation	Sum of Square SS	Degrees of Freedom (DF)	Mean Sum of Squares (MS)	F-Ratio	P-value
Regression	129.018	3	43.006	30.46	0.0000
Error	70.595	50	1.412		
Total	199.613	53			

F-ratio of 30.46 (P-value = 0.0000) indicates that the model fits. Hence we may proceed with further analysis to examine the relationships between the responses by the various categories of subjects. Now the expected (mean) LDL levels for subjects whose test scores are critically below the minimum normal LDL level is estimated form Equations 8 and 16 as;

$$5.90 - 4.92 = 0.98$$

The expected LDL levels for subjects whose test scores are either marginally below the minimum normal LDL level and those subjects whose LDL levels are normal are estimated from Equations 9 and 16 as respectively

5.90 – 4.06 = 1.84 and

5.90 – 2.70 = 3.20.

Finally the mean LDL level of subjects whose test results are critically above the maximum normal LDL level is estimated form Eqns 7C and 16 as 5.90. Note that the expected difference in the mean score of subjects whose test results are critically below the minimum normal score and those whose scores are either marginally below the minimum normal score or marginally above the maximum normal score is estimated from Equation 16 as

$$\begin{aligned}(5.90 - 4.92) - (5.90 - 4.06) &= -4.92 + 4.06 \\ &= -0.86\end{aligned}$$

The null hypothesis that the difference between these mean responses is zero (Eqn. 8) is tested using Equation 9 as

$$t = \frac{-4.92 + 4.06}{\sqrt{(0.129 + 0.204 - 2(-0.042))(1.412)}}$$

$$= \frac{-0.86}{0.767} = -1.121$$

(P-value = 0.1324)

Since $|-1.121| = 1.121 < 2.68 = t_{0.995;50}$

We do not reject H_o of Equation 8

Similarly, the null hypothesis that means scores of subjects whose test results are critically below the minimum normal score and those subjects whose test results are normal, H_o: $\beta_1 = \beta_3$ is tested using Equation 11 as

$$t = \frac{b_1 - b_3}{se(b_1 - b_3)}$$

$$= \frac{-4.92 - (-270)}{\sqrt{(0.129 + 0.246 - 2(-0.061))(1.412)}}$$

= 484.4

Therefore for $\alpha = 0{,}01$ we would reject the null hypothesis that the mean LDL level of subjects whose LDL levels are critically below the minimum normal level is equal to the mean LDL level of subjects with normal test scores or results. The null hypotheses that the mean scores of subjects, whose test results are critically below the minimum normal level and those whose tests results are critically above the maximum normal level is tested using equation 17, namely.

$$t = \frac{(5.90 - 4.92) - 5.90}{\sqrt{(0.129)(1.412)}} = \frac{-4.92}{0.430} = 11.442$$

$$(P - value = 0.0000)$$

Showing that the mean LDL levels of subjects whose LDL levels are below the minimum normal level and subjects whose LDL levels are critically above the maximum normal level are highly statistically different. The hypothesis that the mean LDL level of subjects who tested normal is the same as the mean LDL level of subjects whose LDL levels are critically above the maximum normal level (Eqn 14) is tested using Equation 15 as

$$t = \frac{(5.90 - 2.70) - 5.90}{\sqrt{(0.246)(1.412)}} = \frac{2.70}{0.589} = -4.60$$

$$(P - value = 0.0000)$$

4. Summary and Conclusion

We have in this paper presented a statistical method for the analysis of diagnostic screening test results which are recorded as quantitative scores using dummy variable multiple regression techniques. The proposed method developed estimates of the expected test scores for subjects whose test scores are critically below the minimum normal score, those subjects whose test scores are critically above the maximum normal score, those subjects whose test scores are either marginally below the minimum normal score or marginally above the maximum normal score, and for those subjects whose test results or scores are normal. Test statistics are also developed for testing the existence of any significant differences between the expected scores by these various groups.

References

[1]. Barreto, Humberto; Howland, Frank (2005). "Chapter 22: Dummy Dependent Variable Models". Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel. Cambridge University Press. ISBN 0-521-84319-7

[2]. Boyle, R.P. 1970. “Path Analysis and Ordinal Data”. American Journal of Sociology,Vol.47, PP 461-480.

[3]. Draper, N.R.; Smith, H. (1998) Applied Regression Analysis, Wiley. ISBN 0-471-17082-8 (Chapter 14)

[4]. Neter J., Wasserman W. and Kurtner M.H. 1983. “Applied Linear Regression Models”. Richard D. Irwin Inc. pp 329-330

[5]. Oyeka I.C.A. 1993. “Estimating Effects in Ordinal Dummy Variable Regression”. STATISTICA, anna LIII, n.2, pp 262-268.

[6]. Suits, Daniel B. (1957). "Use of Dummy Variables in Regression Equations". Journal of the American Statistical Association 52 (280): 548–551. JSTOR 2281705.