Iraqi Journal of Statistical Science (15) 2009
p.p. [17-30]

Categorical data analysis with Practical Application

Furat Barakat Hayran Al _ Dassy*

Abstract

The main purpose of this study is to find out vectors that can be inserted to the statistical analysis of certain groups of variables which are formed as a result of a certain condition depending on categorical data (i.e. Qualitative variable).

Many statistical models have been discussed to deal with this condition in order to have a discriminate function or to have variables by which we can specify the condition gradually. Depending on the importance of the inserted variable in that condition, a theorem form of this sensitive variable has been derived . Some applications have been also used in this paper.

تحليل المتغيرات المصنفة مع التطبيق العملي

الملخص

الغرض الرئيسي من هذه الدراسة هو ايجاد متجهات يمكن ادخالها في التحليل الاحصائي لعدة مجموعات من المتغيرات التي تكونت نتيجة لحالة معينة معتمدين في ذلك على المتغيرات المصنفة (المتغيرات النوعية) .

وكذلك تم شرح عدة نماذج احصائية للتعامل مع هذه الحالة لكي نحصل على الدالة المميزة أو للحصول على متغيرات نستطيع فيها تمييز الحالة بالتدرج حسب اهمية هذا المتغير الداخل في هذه الحالة تمهيداً لاشتقاق صيغة نظرية لهذه المتغيرات الحسية وهناك تطبيقات تم استخدمها في هذا البحث .

Introduction

This paper will try to explain samples (models) of vector expectations versus models for every response or order pairs individually or multiply . The discrimination between

*Assistant Lecturer/ Science College _Mathematic Department/Duhok University

one another is known as linear logarithm models (see Mahalanobios (1936), Hand (1981), Fisher (1950) and Anderson (1957)[2].

The discriminate studies concerning the sensitive data or qualitative data are limited and Fisher (1950) disregarded and them . This case was treated by random specific values where two common approximates of the prospects equations are compared and which are relates to the assumed statistic knowledge and as follows:

First : Concerning all the knowledge at one point then the estimation of the regression knowledge by statistically .

Secondly : The estimation of the regression and the united statistical knowledge is done at one time .

What we are doing ,in particular, is discriminating the independence of each response for the other illustrative variables or describing the union among responses (i.e Qualitative variables).

In general ,our style (method) of response analysis depends on applying the formula of each option , response or class of variables as value for each item with two basic ideas of the discriminate analysis of the quantitative data and of the discriminate analysis of the paired data .It is possible first to determine the classes characteristics transfer from quantitative data into qualitative data and then to be used in the discriminate analysis in order to find out the variables and their importance for a case like the quantitative data . The work of Tielschetal (1989) is considered as an example for this , He conducted a theoretical analysis for the specific data of the vision power that was taken from a survey mode on 5000 persons at age of 40 . This sample was tested as it represent the total population from which it had been randomly taken.

This work is to determine the quantitative data that affect the eye power then observing it until the examination result . Its importance will be according to the examination result and

it will be distinguished in the same way as the power of the level of education, taking medicine and its quantity , … etc in which the versus regression model clarifies the importance of the middle versus through the variables .

weighted sum of squares [4];

The identity matrix is a special case of a diagonal matrix . This matrix arises in a number of situation .we minimize a weighted sum of squares

$$s(\beta)=\sum_{i=1}^{n} w_i (y_i - x'_i \beta)^2,$$

Where w_i is the value of the weight for the *ith* observation the vector w would represent the $n\times 1$ vector of weights .if we define w to be the $n\times n$ matrix with w_i on the main diagonal ,such that

$$w=\begin{pmatrix} w1 & & & 0 \\ & w2 & & \\ & & . & \\ & & . & \\ 0 & & & wn \end{pmatrix},$$

We can write the weighted sum of squares function in matrix notation as

$$s(\beta)=(y-x\beta)'\ w(y-x\beta).$$

given that regression model in matrix form is :-

$$y=x\beta+\varepsilon,$$

The OLS estimator of β is found by minimizing the sum of squares

$$\begin{aligned} s(\beta) &= (y-x\beta)'\ (y-x\beta) \\ &= y\ y-2y'\ x\beta+\beta'\ x'\ x\beta. \end{aligned}$$

The vector of partial derivatives of the sum of squares with respect to β may be expressed as

$$\frac{\partial s(\beta)}{\partial \beta} = -2x'\ y + 2x'\ x\beta.$$

When we equate this expression to zero and solve , it yields the OLS estimator

$$b = (x'\ x)^{-1} x'\ y,$$

Where b is the $(k+1)\times 1$ vector of OLS estimates. The variance –covariance matrix of b is obtained by multiplying the mean square error , σ_{ε}^{2}, or its estimate ,by the inverse of the sum of squares and cross-products matrix , $x'x$, as follows : $\hat{\text{var}}(b) = \partial_{\hat{\varepsilon}}^{2}(x'x)^{-1} = w^{-1}$

This operation results in a square matrix with $\text{var}(b_k)$ along the diagonal and $\text{cov}(b_k, (j \neq k)$ on the off-diagonal .

Binary variables :

The scientists kuder and Richardson at (1933)[5] represent the first way which used the alternative variables which is called by (binary scores) .Which is the responses mean which depends on one another , when we ask about a specific phenomenon the respond will be yes or no .which win give it a specific value like (1 , 0) and $\left(\frac{1}{3}, \frac{2}{3}\right)$ in which there is a hard difficulty for reaching to

the like values which reduces the dissimilarity inside the units (between the testing) At 1935 (Horst) [6] mentioned that the selection will adjust to give the greatest discriminate value between it, and he depended on the deaf-variables ,

which we refuse it recently ,and which it contain variables capable of sensory classifying to specific numbers and this is which it increases its difficulty in the discriminate analysis by depending on the numbers responses and acts three variables Y_1, Y_2, Y_3 and three responses X_1, X_2, X_3 it has a specific frequencies like $f_{11}, f_{12}, \ldots, f_{33}$ in which we can form this common table :

	Y_1	Y_2	Y_3		
Y_1	f_{11}	f_{12}	f_{13}	$\sum f_{ij}$	$Y_1 = \sum x_j f_{1j}$
Y_2	f_{21}	f_{22}	f_{23}	$\sum f_{2j}$	$Y_2 = \sum x_j f_{2j}$
Y_3	f_{31}	f_{32}	f_{33}	$\sum f_{3j}$	$Y_3 = \sum x_j f_{3j}$
Total	$\sum f_{i1}$	$\sum f_{i2}$	$\sum f_{i3}$		

$$f_t = \sum\sum f_{ij} \quad , \quad c = \frac{y_t^2}{f_t} = \frac{\left(\sum\sum x_{ij} f_{ij}\right)^2}{f_t}$$

$$Y_t = \sum_j \sum_i x_j f_{ij}$$

$$= x_1 f_{11} + x_2 f_{12} + x_3 f_{13} + x_1 f_{21} + x_2 f_{22} + x_1 f_{31} + x_2 f_{32} + x_3 f_{33}$$

$$\sum\sum Y_{ij^2} = \sum\sum x_{ij} f_{ij}$$

Which we can represent X_{ij} by X_i^2 or X_j^2 which is equal to X_1^2, X_2^2, X_3^2 respectively

$$F = \begin{pmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{pmatrix} \quad , \quad f = \begin{pmatrix} \sum f_{i1} \\ \sum f_{i2} \\ \sum f_{i3} \end{pmatrix}$$

$$D = \begin{pmatrix} \sum f_{i1} & 0 & 0 \\ 0 & \sum f_{i2} & 0 \\ 0 & 0 & \sum f_{i3} \end{pmatrix} \quad ,$$

$$D = \begin{pmatrix} \sum f_{1j} & 0 & 0 \\ 0 & \sum f_{2j} & 0 \\ 0 & 0 & \sum f_{3j} \end{pmatrix}$$

$$\underline{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad ,$$

$$Y_t^2 / f_t = X' \left(\frac{f\ f'}{f_t} \right) X$$

$$\sum\sum Y_{ij}^2 = X'DY$$

Hence

$$sst = \sum\sum Y_{ij}^2 - Y_t^2 / f_t = \underline{X}'DX - \underline{X}'(ff'/f_t)\underline{X}$$

$$= \underline{X}'(D - ff'/f_t)\underline{X}$$

$$ssb = \sum Y_j^2 / f_j - Y_t^2 / f_t$$

$$ssb = \underline{X}' \left(F'D_n^{-1}F - ff'/f_t \right) \underline{X}$$

$$F'D_n^{-1}F = \begin{bmatrix} \sum f_{i1}^2 / \sum\sum f_{ij} & \sum f_{i1} f_{i2} / \sum\sum f_{ij} & \sum f_{i1} f_{i3} / \sum\sum f_{ij} \\ \sum f_{i1} f_{i2} / \sum\sum f_{ij} & \sum f_{i2}^2 / \sum\sum f_{ij} & \sum f_{i2} f_{i3} / \sum\sum f_{ij} \\ \sum f_{i1} f_{i3} / \sum\sum f_{ij} & \sum f_{i2} f_{i3} / \sum\sum f_{ij} & \sum f_{i3}^2 / \sum\sum f_{ij} \end{bmatrix}$$

$$ff'/f_t = \begin{bmatrix} (\sum f_{i1})^2/\sum\sum f_{ij} & \sum f_{i1}\sum f_{i2}/\sum\sum f_{ij} & \sum f_{i3}\sum f_{i1}/\sum\sum f_{ij} \\ \sum f_{i1}f_{i2}/\sum\sum f_{ij} & (\sum f_{i2})^2/\sum\sum f_{ij} & \sum f_{i3}\sum f_{i2}/\sum\sum f_{ij} \\ \sum f_{i1}\sum f_{i3}/\sum\sum f_{ij} & \sum f_{i2}\sum f_{i3}/\sum\sum f_{ij} & (\sum f_{i3})^2/\sum\sum f_{ij} \end{bmatrix}$$

so

$$F'D_n^{-1}F - ff'/f_t = \begin{bmatrix} \dfrac{\sum f_{i1}^2 - (\sum f_{i1})^2}{\sum\sum f_{ij}} & \dfrac{\sum f_{i1}f_{i2} - \sum f_{i1}\sum f_{i2}}{\sum\sum f_{ij}} & \dfrac{\sum f_{i1}f_{i3} - \sum f_{i1}\sum f_{i3}}{\sum\sum f_{ij}} \\ 0 & \dfrac{\sum f_{i2}^2 - (\sum f_{i2})^2}{\sum\sum f_{ij}} & \dfrac{\sum f_{i2}f_{i3} - \sum f_{i2}\sum f_{i3}}{\sum\sum f_{ij}} \\ 0 & 0 & \dfrac{\sum f_{i3}^2 - (\sum f_{i3})^2}{\sum\sum f_{ij}} \end{bmatrix}$$

And also it will be

$$\frac{ssb}{sst} = \frac{X'\left(F'D_n^{-1}F - f\ f'/f_t\right)X}{X'\left(D - f\ f'/f_t\right)X}$$

By using Lacherange multiplier [1] and statistics and athematics processes the researchers have all the detail about it in getting the matrix .

$$C_1 = D^{-\frac{1}{2}} F' D_n^{-1} F D^{-\frac{1}{2}} - \eta^2 WW'/W'W$$

$$\eta^2 = 1 \qquad \& \qquad W = D^{\frac{1}{2}} I$$

Which the other matrix are represented simply by C_1

$$C_1 = D^{-\frac{1}{2}} F' D_n^{-1} F D^{-\frac{1}{2}} - (1/f_t) D^{\frac{1}{2}} II' D^{\frac{1}{2}}$$

We follow the frequency process in which we multiply C_1 by any arbitrariness vector as $b_0 = \begin{bmatrix} +1 \\ 0 \\ -1 \end{bmatrix}$, in which the sum of its elements equal zero and we get $b_1 = C_1 b_0$ then we divide the elements of b_1 over the bigger element of the

b_1 elements which it represented by $|k_1|$ and the result represented by b_1^* which

$$\frac{b_1}{|k_1|} = b_1^*$$

Thus

$$\frac{b_{j+1}}{|k_{j+1}|} = b_{j+1}^*$$

Until we get

$$b_j^* = b_{j+1}$$

Using the b_{j+1} as sample vector which is regarded the higher sample value of

C_1 which it gives by $|k_{j+1}|$

$$W = \left(f_t / b'^{*}_{j+1} b_{j+1}\right)^{\frac{1}{2}} b^{*}_{j+1}$$

$$W'W = f_t$$

$$\underline{X} = D^{-\frac{1}{2}} w$$

$$\therefore Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \frac{D_n^{-1} F \underline{X}}{\eta}$$

The variable which carries highest value is the best for the discriminate process.

Following the variable carries the lowest value ,and follow for any variables .

Practical Application:

One of the factories announced its requirement for a huge number of personnel, and that which carries different degrees to work at that factory ,which had been carried out by one of the architecture companies .and the difficult that faced this factory was that the number of personnel that applied for a job were above the factory's requirement .

Also it was hard to determine the personnel's skill and capability for the job and statistics features had been putting in the effect in which the stipulation of age and testimony had been putting aside ,and been settled previously the statistics authorities noticed by many questionnaires the following .

The features that were affected and are affecting the architecture process course and experimental working that had been continuous for a year ,it was as following :-

1- sex male female
2- housing situation very good good
medium bad
3- final education average very good good
medium bad
4- nearness of work center very good good
medium bad
5- family income very good good
medium bad
6- range of caring of general tracking very good good
medium bad

These things were the causing that were influencing the work path and acceptance of the project works .

So ,we want to apply these variable s to distinguish the state for every worker before his acceptance in this factory .and the binary discriminate data were the excellent directions here .

And for simple analysis ,male and female have been analyzed one by one .And a100 item has been taken from each sex as random sample , representation the total population to obtain discriminate variable to get benefit from them in the new acceptance ,and for easiness the last two tests have been mixed and formed that table bellow:

	males			female		
	very good	good	medium	very good	good	medium
housing situation	60	30	10	20	60	20
final education average	40	25	35	15	30	55
nearness of work center	55	20	25	33	10	57
family income	20	30	50	10	20	70
Range of caring of general tracking	25	25	50	70	20	10
total	200	130	170	148	140	212

About the males we will get

$$Y = \frac{D_n^{-1} F \underline{X}}{\eta} = \begin{bmatrix} 1.4568 \\ 0.4562 \\ 0.998 \\ 1.3426 \\ 0.89321 \end{bmatrix}$$

We observed the variables according to their importance .

1. housing situation ,in first class that affects on the work path .
2. family income comes at the second class .
3. the nearness and the farness come at the third class.
4. the elegance and the care ness of appearances at the penultimate .
5. finally ,the average of the final education average .

In the case when any male advances to work at this factory it is enough to give him this list to fill it and then we can expect if he will be a good works or not hence we will give the superior checker a recommendation to decide the acceptance or the refusal .

As for females , the results were as following :

$$Y = \frac{D_n^{-1} F \underline{X}}{\eta} = \begin{bmatrix} 1.3333 \\ 1.4321 \\ 0.8899 \\ 1.7894 \\ 0.9899 \end{bmatrix}$$

Which it was :

1. family 's income comes at the first class.
2. professional education average comes at the second class .
3. housing situation at the third class.
4. care ness of the elegance at the penultimate ,and in this the males and females are equal .
5. finally , the nearness and the farness of the work center .

The designation is obeys the central designation plan in the country. For both male and female .

Conclusion and recommendation

We conclude from what we displayed that there is a statistical method which
had been neglected by many statistics application studies, and it is the ability of

getting benefit from the qualitative data and classification . Then distinguish its
importance in affecting a specific phenomenon .

We advise for increasing the statistics researches and then using them on the real life for getting benefit from it as we showed in our research .We derive and take a specific method for treating with qualitative data by depending on various resources.

References

1. Andersen, p.k.(1982)," testing goodness of fit of cox's regression and life model". Biometric, 38, 67-77.
2. Anderson, J. A. and sethilselvan, A. (1982),"a two-step regression model for hazard function". Appl. Stutists. 31-44-51.
3. Bartholomew D. V. (2003)," A two-sample censored data rank test for acceleration" biometrics, 40, 1049-1062.
4. Daniel A .P (1999) ,"statistical methods for categorical data analysis" . university of Texas at Austin .academic press, Inc.
5. Morean, T. O. Qaigleg, J. abd Lellouch, J. (2003) on" D. Schoen field's approach for testing the proportional hazard assumption". Biometrik, 73, 513-515.
6. Shizuhiko Nishisato, (1980)," Analysis of categorical data"; university of Toronto press buffalo London, dual scaling & its applications.