Detection of Dependence Between

Two Groups of Variables

A. H. Pooi

Sunway University Business School
Sunway University
Bandar Sunway, Malaysia

Abstract

Let $F_1$, $F_2$ and $F_3$ be three groups of variables some of which may be categorical. A method is proposed to determine whether there is any dependence between the first two groups $F_1$ and $F_2$ of variables. After establishing the dependence between $F_1$ and $F_2$, another method is proposed to determine whether $F_2$ is dependent on $F_3$ in the presence of $F_1$. When the tests are applied to logistic and polychotomous regression examples, they are found to have good discerning ability when there is a dependence between the two groups under study.

Mathematics Subject Classification: 62H15, 62H17

Keywords: Binary variables, Multivariate power normal mixture distribution, Resampling

1. Introduction

Very often the data are collected by categories. Categorical variables may be classified as nominal or ordinal. The categories associated with a nominal variable have no particular order while those associated with an ordinal variable have some inherent ordering.
Contingency table analysis ([1]–[5]), logistic regression ([6]) and polychotomous regression ([7]–[8]) are some examples in which we try to investigate the dependence among the variables of which some may be categorical.

This paper is an attempt to detect dependence among the variables some of which may be categorical.

Consider the situation when we have three groups of variables:

\[
F_1: y_{11}', y_{12}', \ldots, y_{1m}' \\
F_2: y_{21}', y_{22}', \ldots, y_{2n}' \\
F_3: y_{1m+1}', y_{1m+2}', \ldots, y_{1m+p}'
\]

which may possibly include some categorical variables.

A categorical variable \( y_{ij} \) of \( n_c \) categories may first be coded by using a vector \( (y_{ij1}, y_{ij2}, \ldots, y_{ijn_c}) \) of \( n_c - 1 \) binary variables. For example, for \( 1 \leq i \leq n_c - 1 \), the \( i \)-th category may be coded as \( (0, 0, \ldots, 1, 0, \ldots, 0) \) in which the value 1 appears as the \( i \)-th entry, while the \( n_c \)-th category may be coded as a vector of zeros. After replacing all the categorical variables by the corresponding binary variables, the three groups of variables may be expressed as

\[
G_1: y_{11}, y_{12}, \ldots, y_{1m} \\
G_2: y_{21}, y_{22}, \ldots, y_{2n} \\
G_3: y_{1m+1}, y_{1m+2}, \ldots, y_{1m+p}
\]

where each \( y_{ij} \) may originally be equal to \( y_{ij1}' \) or \( y_{ij2}' \) for some \( j_1 \) or \( j_2 \).

We may be interested in examining the dependence between \( G_1 \) and \( G_2 \). For example in a situation in which we are interested in the factors that influence whether a political candidate wins an election, \( G_1 \) may contain

\[
y_{11} = \text{amount of money spent on the campaign} \\
y_{12} = \begin{cases} 
  1 & \text{if the candidate is an incumbent} \\
  0 & \text{if the candidate is not an incumbent}
\end{cases}
\]

and \( G_2 \) may contain

\[
y_{21} = \begin{cases} 
  1 & \text{if the candidate wins the election} \\
  0 & \text{if the candidate loses the election}
\end{cases}
\]
After establishing the dependence between $G_1$ and $G_2$, we may be interested to find out whether the third group $G_3$ which contains the variable $y_{13} = \text{amount of time spent on the campaign}$ will further influence the results of the election. This is then a situation in which we wish to determine the dependence of $G_2$ on $G_3$ in the presence of $G_1$.

The layout of the paper is as follows. In Section 2, a new distribution, called the multivariate power-normal mixture distribution, is introduced. The new distribution forms the foundation for the development of the paper. In Section 3, a method is proposed to determine the dependence between $G_1$ and $G_2$. Section 4 is devoted to the description of the method for detecting the dependence between $G_2$ and $G_3$ in the presence of $G_1$. In Section 5, an example on logistic regression is used to illustrate the method in Section 3, and the performance of the method is also compared with that based on deviance. In Section 6, an example on polychotomous regression is used to illustrate the method in Section 4, and the performance as given by the power of the test is given. Section 7 concludes the paper.

2. Multivariate Power-normal Mixture Distribution

Yeo and Johnson [9] proposed the following power transformation:

$$
\tilde{z} = \psi(\lambda^+, \lambda^-, z) = \begin{cases} 
(z+1)^{\lambda^+} -1/\lambda^-, & z \geq 0, \lambda^- \neq 0 \\
\log(z+1), & z \geq 0, \lambda^- = 0 \\
-(z-1)^{\lambda^-} -1/\lambda^+, & z < 0, \lambda^+ \neq 0 \\
-\log(-z+1), & z < 0, \lambda^+ = 0 
\end{cases}
$$

(1)

If $z$ has the standard normal distribution, then $\tilde{z}$ has a non-normal distribution which is derived by the power transformation. The distribution of $\tilde{z}$ may be called the power-normal distribution.

Next, consider a vector $y$ consisting of $k$ correlated random variables. The vector $y$ is said to have a $k$-dimensional power-normal distribution [10] with parameters $\mu, H, \lambda_i^+, \lambda_i^-, \sigma_i, 1 \leq i \leq k$ if

$$
y = \mu + H \varepsilon
$$

(2)
where $\mathbf{y} = E(y)$, $H$ is an orthogonal matrix, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$ are uncorrelated,

$$
\varepsilon_i = \sigma_i \left[ \tilde{\varepsilon}_i - E(\tilde{\varepsilon}_i) \right] \sqrt{\text{var}(\tilde{\varepsilon}_i)}^{1/2}
$$

(3)

$\sigma > 0$ is a constant, and $\tilde{\varepsilon}_i$ has a power-normal distribution with parameters $\lambda^+$ and $\lambda^-$. The above power-normal distribution may be extended to the following power-normal mixture distribution.

Suppose $\tilde{\varepsilon}_i^{(j)}$ has a power-normal distribution with parameters $\lambda_i^{(j)+}$ and $\lambda_i^{(j)-}$, $j = 1, 2$. For $0 \leq p_i \leq 1$, define

$$
\tilde{\varepsilon}_i^* = \begin{cases} 
\tilde{\varepsilon}_i^{(1)} - \mu_{i1} & \text{with probability } p_i \\
\tilde{\varepsilon}_i^{(2)} + \mu_{i2} & \text{with probability } q_i = 1 - p_i 
\end{cases}
$$

where $\mu_{i1}$ and $\mu_{i2}$ are constants,

$$
\tilde{\varepsilon}_i^{(j)} = \frac{\tilde{\varepsilon}_i^{(j)} - E(\tilde{\varepsilon}_i^{(j)})}{\sqrt{\text{var}(\tilde{\varepsilon}_i^{(j)})}}, \quad j = 1, 2
$$

and $E(\tilde{\varepsilon}_i^*) = 0$ (4)

In order to achieve the condition given by Equation (4), the constant $\mu_{i2}$ should be given by

$$
\mu_{i2} = \frac{p_i \mu_{i1}}{q_i}
$$

The variance of $\tilde{\varepsilon}_i^*$ is then given by

$$
\text{var}(\tilde{\varepsilon}_i^*) = 1 + \left( p_i + p_i^2 / q_i \right) \mu_{i1}^2.
$$

We may refer to the random variable $\tilde{\varepsilon}_i^*$ as one which has a power-normal mixture distribution with parameters $p_i$, $\mu_{i1}$, $\lambda_i^{(j)+}$ and $\lambda_i^{(j)-}, j = 1, 2$.

We next define

$$
\varepsilon_i^* = \begin{cases} 
\sigma_i^+ \left( \tilde{\varepsilon}_i^{(1)} - \mu_{i1} \right) \sqrt{\text{var}(\tilde{\varepsilon}_i^*)} & \text{with probability } p_i \\
\sigma_i^- \left( \tilde{\varepsilon}_i^{(2)} + \mu_{i2} \right) \sqrt{\text{var}(\tilde{\varepsilon}_i^*)} & \text{with probability } q_i 
\end{cases}
$$

(5)
Now consider a vector $y^*$ consisting of $k$ correlated random variables. The vector $y^*$ is said to have a $k$-dimensional power-normal mixture distribution with parameters $\mu^*, H^*, p_i, \mu_j^*, \sigma_j^*, \lambda_i^{(j)*}, \lambda_j^{(j)*}, 1 \leq i \leq k, 1 \leq j \leq 2$ if

$$y^* = \mu^* + H^* \epsilon^*$$  \hspace{1cm} (6)

where $\mu^* = E(y^*)$; $H^*$ is an orthogonal matrix, $\epsilon_1^*, \epsilon_2^*, \ldots, \epsilon_k^*$ are uncorrelated and $\epsilon_i^*$ is given by Equation (5).

3. Detection of Dependence Between Two Groups

The observations on the first group of variables and the $j$-th variable in the second group may first be fitted with an $(m+1)$-dimensional power-normal mixture distribution using the method based on moments ([10]).

For each set of observed values in the first group, a conditional distribution for the $j$-th variable in $G_2$ may be obtained using the numerical method given in [10]. The mean of the conditional distribution may then be regressed on the variables in $G_1$ to get the estimated values $\hat{\beta}_{j1}, \hat{\beta}_{j2}, \ldots, \hat{\beta}_{jm}$ for the slope parameters.

Let $\hat{\beta}$ be the vector formed by all such estimated coefficients:

$$\hat{\beta}^T = (\hat{\beta}_{11}, \hat{\beta}_{12}, \ldots, \hat{\beta}_{1m}, \hat{\beta}_{21}, \hat{\beta}_{22}, \ldots, \hat{\beta}_{2m}, \ldots, \hat{\beta}_{n1}, \hat{\beta}_{m1}, \ldots, \hat{\beta}_{mn})$$  \hspace{1cm} (7)

Denoting $\alpha$ as a small chosen probability, we next we find a region $R_\alpha$ in the $nm$-dimensional space such that $R_\alpha$ will include $\hat{\beta}$ with an approximate probability of $1-\alpha$ if there is no dependence between $G_1$ and $G_2$.

Suppose $N$ is the number of sets of observed values of $y_{11}, y_{12}, \ldots, y_{1m}, y_{21}, y_{22}, \ldots, y_{2n}$. The steps for finding the region $R_\alpha$ are as follows:

1. Choose at random $N$ integers without replacement from $\{1, 2, \ldots, N\}$.
   Let the chosen integers be $j_1, j_2, \ldots, j_N$.

2. Form a table of $N$ rows of which the $i$-th row is formed from the $i$-th observed values of $y_{1i}, y_{12}, \ldots, y_{1m}$ and the $j_i$-th observed values of $y_{21}, y_{22}, \ldots, y_{2n}$.
3. Fit the data given by the first \( m \) columns, and the \((m+j)\)-th column of the table in 2, with an \((m+1)\)-dimensional power-normal mixture distribution.

4. For each set of observed values of the variables in the first group, a conditional distribution for the \( j \)-th variable in \( G_2 \) is obtained by using the multivariate power-normal mixture distribution in 3.

5. The mean of the conditional distribution in 4 is then regressed on the variables in \( G_1 \) to get the estimated values \( \tilde{\beta}_{j1}, \tilde{\beta}_{j2}, \ldots, \tilde{\beta}_{jm} \) of the slope parameters. Let \( \tilde{\beta} \) be the vector formed by all such coefficients:
   \[
   \tilde{\beta}^T = (\tilde{\beta}_{11}, \tilde{\beta}_{12}, \ldots, \tilde{\beta}_{1m}, \tilde{\beta}_{21}, \tilde{\beta}_{22}, \ldots, \tilde{\beta}_{2m}, \ldots, \tilde{\beta}_{n1}, \tilde{\beta}_{n2}, \ldots, \tilde{\beta}_{nm})
   \]

6. Repeat Steps 1–5 \( N_2 \) times and fit the \( N_2 \) generated values of \( \tilde{\beta} \) with an \( nm \)-dimensional power-normal distribution.

7. The random variable \( \tilde{\beta} \) may be written as
   \[
   \tilde{\beta} = R_\beta + H_\beta \varepsilon_\beta
   \]
of which the right side has a structure which is similar to that of (2). Thus like \( \varepsilon_i \) in Equation (3), \( \varepsilon_\beta \) is a function of a random variable (denoted as \( z_\beta \)) which has a standard normal distribution. A candidate for \( R_\alpha \) may be formed from the values of \( \tilde{\beta} \) of which the corresponding \( z_{\beta 1}, z_{\beta 2}, \ldots, z_{\beta \text{nm}} \) have a sum of squares
   \[
   D^2 = \sum_{i=1}^{nm} z_{\beta i}^2
   \]
which is less than or equal to the \( 100(1-\alpha) \)% point \( \chi^2_{nm, \alpha} \) of the chi-square distribution with \( nm \) degrees of freedom.

To test \( H_0 \): There is no dependence between \( G_1 \) and \( G_2 \), we may adopt the following decision rule:

“Accept \( H_0 \) if \( \hat{\beta} \in R_\alpha \).”

4. Detection of Dependence Between Two Groups in the Presence of Another Group

After establishing the dependence between the groups \( G_1 \) and \( G_2 \), we may wish to examine whether another group \( G_3 \) will influence \( G_2 \) further after \( G_1 \) has exerted its influence on \( G_2 \). A method for examining the dependence between \( G_2 \) and \( G_3 \) in the presence of \( G_1 \) is as follows.
Detection of dependence between two groups of variables

The observations on the first group of variables and the $j$-th variable in the second group may first be fitted with an $(m+1)$-dimensional power-normal mixture distribution.

For each set of observed values in the first group, a conditional distribution for the $j$-th variable in $G_2$ may be obtained. The mean of the conditional distribution may be subtracted from the observed value of the $j$-th variable to form a residual. A total of $N$ such residuals can be computed.

The computed residuals may be regarded as observed values of a certain random variable denoted as $r_{2j}$. The method in Section 3 is next used to examine the dependence between $G_3$ and the group $G_r$ formed by the variables $r_{21}, r_{22}, \ldots, r_{2n}$. If there is no dependence between $G_3$ and $G_r$, then we conclude that there is no dependence between $G_2$ and $G_3$ in the presence of $G_1$.

5. Logistic Regression Example

Let $G_1$ be the group which contains the variable $y_{11}$ which has a standard normal distribution and $G_2$ the group which is made up of the variable $y_{21}$ given by

$$y_{21} = \begin{cases} 1 & \text{with probability } 1/(1 + e^{-(\beta_0 + \beta_1 y_{11})}) \\ 0 & \text{with probability } 1 - 1/(1 + e^{-(\beta_0 + \beta_1 y_{11})}) \end{cases}$$

where $\beta_0$ and $\beta_1$ are constants.

Then there is no dependence between the groups $G_1$ and $G_2$ if $\beta_1 = 0$. On the other hand, there will be a dependence between $G_1$ and $G_2$ if $\beta_1 \neq 0$, and the dependence will be stronger if $|\beta_1|$ gets larger.

We may use the following procedure to evaluate the performance of the test in Section 3:

a) Generate $N$ values of $(y_{11}, y_{21})$ and compute $\hat{\beta}$ (see Equation (7)).
b) Use the method in Section 3 to find $N_2$ values of $\tilde{\beta}$.
c) Find $R_a$ using the $N_2$ values of $\hat{\beta}$ in Step b).

d) Determine whether $\hat{\beta} \in R_a$.

e) Repeat Steps a)-d) for $N_1$ times and use the proportion of times $\hat{\beta} \not\in R_a$ to estimate the power of the test.

Table 1 shows the estimated powers of the proposed test and the test based on deviance.

<table>
<thead>
<tr>
<th>$\beta_i$</th>
<th>Test based on deviance ($N_1=1000$)</th>
<th>Proposed test ($N_1 = N_2=100$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.051</td>
<td>0.04</td>
</tr>
<tr>
<td>0.5</td>
<td>0.460</td>
<td>0.53</td>
</tr>
<tr>
<td>1.0</td>
<td>0.928</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Table 1 shows that the powers of the proposed test are comparable to those of the test based on deviance.

6. Polychotomous Regression Example

Let $G_1$ be the group which contains the variables $y_{11}$ and $y_{12}$ which are independent and normally distribution with mean 0 and variance 1, and $G_2$ the group which is made up of the variables $y_{21}$ and $y_{22}$ given by

$$(y_{21}, y_{22}) = \begin{cases} 
(1, 0) & \text{with probability } e^{\beta_1^{(1)} y_{11} + \beta_1^{(2)} y_{12}} / P_T \\
(0, 1) & \text{with probability } e^{\beta_2^{(1)} y_{11} + \beta_2^{(2)} y_{12}} / P_T \\
(0, 0) & \text{with probability } 1 / P_T 
\end{cases}$$

where $P_T = (1 + e^{\beta_1^{(1)} y_{11} + \beta_1^{(2)} y_{12}} + e^{\beta_2^{(1)} y_{11} + \beta_2^{(2)} y_{12}})$.

Then there is no dependence between the groups $G_1$ and $G_2$ if $\beta_1^{(1)} = \beta_1^{(2)} = 0$. On the other hand, there will be a dependence between $G_1$ and $G_2$ if $\beta_1^{(1)} \neq \beta_1^{(2)}$ or $\beta_2^{(1)} = 0$, and the dependence will be stronger if $|\beta_1^{(1)}|$ or $|\beta_1^{(2)}|$ gets larger.

To evaluate the performance of the test for the dependence between $G_1$ and $G_2$ in Section 3, we may perform the steps which are similar to Steps a)–d) in Section 5.
Table 2 shows the power of the test for the dependence between $G_1$ and $G_2$.

Table 2: Power of the test in the polychotomous regression example ($\beta_0^{(1)} = 0.30$, $\beta_0^{(2)} = 0.25$, $\alpha=0.05$, $N=60$, $N_1=N_2=100$).

<table>
<thead>
<tr>
<th>$\beta_1^{(1)}$</th>
<th>$\beta_1^{(2)}$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.56</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table 2 shows that the power under the null hypothesis of no dependence is close to the value $\alpha=0.05$, and when there is a departure from the null hypothesis, the power increases.

To evaluate the performance of the test in Section 4, we let $G_1 = \{y_{11}\}$, $G_2 = \{y_{21}, y_{22}\}$, $G_3 = \{y_{12}\}$ and perform the following steps:

(i) Generate $N$ values of $(y_{11}, y_{21}, y_{22})$.
(ii) Compute the values of the residuals $r_{21}$ and $r_{22}$.
(iii) Use the method in Section 3 to examine whether there is a dependence between $G_3$ and the group $G_i$ formed by $r_{21}$ and $r_{22}$.
(iv) Repeat Steps (i)–(iii) for $N_i$ times and use the proportion of times the conclusion in (iii) is negative to estimate the power of the test.

Table 3 shows the power of the test for the dependence between $G_2$ and $G_3$ in the presence of $G_1$.

Table 3: Power of the test for dependence between $G_2$ and $G_3$ in the presence of $G_1$ ($\beta_0^{(1)} = 0.30$, $\beta_1^{(1)} = 1.0$, $\beta_0^{(2)} = 0.25$, $\alpha=0.05$, $N=60$, $N_1=N_2=100$).

<table>
<thead>
<tr>
<th>$\beta_1^{(2)}$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0.5</td>
<td>0.34</td>
</tr>
<tr>
<td>1.0</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 3 shows that the test for dependence between $G_2$ and $G_3$ in the presence of $G_1$ has good discerning ability when $\beta_1^{(2)}$ is large. However when $\beta_1^{(2)}$ is zero, the power of 0.10 is somewhat larger than the target value $\alpha=0.05$. Future research may be carried out to investigate further the power of the test.
under the null hypothesis, and perform a refinement of the test if that is deemed necessary.

7. Conclusion

The tests introduced in this paper may be applied to the data which contain values for categorical variables. The tests have good discerning ability when there is a deviation from dependence between the two groups under study. It should be interesting to compare the proposed tests with the existing tests in the literature, in particular those used in contingency table analysis, linear regression models, generalized linear models and other categorical data analysis.

References


[5] K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philosophical Magazine, Series 5, 50(1900), 157–175.


Detection of dependence between two groups of variables


Received: July 15, 2014