Journal of Business and Economics, ISSN 2155-7950, USA
December 2013, Volume 4, No. 12, pp. 1230-1235

http://www.academicstar.us

Elements of Modeling with Mixed Variables

Farrokh Guiahi

(Department of IT/QM, Zarb School of Business, Hofstra University, Hempstead NY 11549, USA)

Abstract: This paper discusses briefly models for mixed variables. In particular, suitable joint distribution of mixed variables is provided by reference to specific examples. Procedures based on model comparison are utilized to study the dependency structure pertaining to a categorical and a continuous variable. Estimation of parameters and computation of Likelihood function is addressed by providing the necessary code as an ADMB program.

Key words: models for mixed variables; ADMB program

JEL code: C18

1. Introduction

This paper highlights some aspects of models for mixed variables commonly encountered in data mining, but less emphasized in basic statistics courses. There are two points worth emphasizing here. First, variables measured on different scales occur more commonly in large data sets. When categorical and continuous variables are studied jointly, more specialized models are needed to accommodate these situations. Apart from modeling issues, segmentation of data arising in cluster analysis, requires special treatment of similarity metrics in the case of mixed variables.

Second, data mining projects typically consider many variables. Due to this high-dimensionally aspect of the data sets, the specification of suitable joint multivariate distribution function for mixed variables becomes a more challenging task.

Section 2, below, discusses log-linear models suitable for count (frequency) data. By reference to an example, we consider the alternative types of dependency for the variables considered, and illustrate schematically the dependency structure by using graphs. Section 3 considers briefly multivariate normal distribution, the classical model for the distribution of a finite set of continuous variables. Distribution for mixed variables is discussed in section 4 with reference to a simple example. Furthermore, we discuss parameter estimation and model selection procedures based on likelihood principle. A program using ADMB code is provided for numerical solution to the likelihood estimation for our example. Some concluding remarks are made in section 5.

2. Categorical Variables

Salient information about categorical variables is summarized by a Table of Counts. The appropriate joint distribution of the variables is a multinomial distribution. The model of interest for discussing relationship among the variables is the log-linear model. The interested reader may refer to Alan Agresti (2002), Stephen E. Fienberg

FarrokhGuiahi, Ph.D., Associate Professor, Zarb School of Business, Hofstra University; research areas: applied statistics, data mining, time series analysis. E-mail: Farrokh.guiahi@hofstra.edu.

(1980) or Ronald Christensen (1997) for further information regarding the analysis of categorical variables.

Categorical variables are labeled as A, B, C, etc. Let us consider an example involving three categorical variables A, B, and C taking values in respective sets {1,2}, {1,2,3} and {1,2}. Information about these variables is summarized by Tables of Counts and Probabilities as follows:

Table 1 Counts & Probabilities When C = 1

n_{111}	n_{121}	n_{131}	p_{111}	p_{121}	p_{131}
n_{111}	n_{121}	n_{131}	p_{111}	p_{121}	p_{131}

Table 2 Counts & Probabilities When C = 2

n_{112}	n_{122}	n_{132}	p_{112}	p_{122}	p_{132}
n_{112}	n_{122}	n_{132}	p_{112}	p_{122}	p_{132}

The cell counts n_{jkl} j = 1,2, k = 1,2,3; l = 1,2 are observed, but cell probabilities p_{jkl} j = 1,2, k = 1,2,3; l = 1,2 are unknown parameters. These probabilities need to be estimated from data. Furthermore, the expected cell count is denoted by $m_{jkl} = Np_{jkl,}$ where $N = \sum_{j,k,l} n_{j,k,l}$ We shall use p to denote the number of categorical variables. In the above example p = 3.

A log-linear model for the above example is given by

$$\log(m_{jkl}) = u + u_j^A + u_k^B + u_l^C + u_{jk}^{AB} + u_{jl}^{AC} + u_{kl}^{BC} + u_{jkl}^{ABC} \tag{1}$$

In Equation (1) above, the components $u_{jk}^{AB}, u_{jl}^{AC}, u_{kl}^{BC}, and\ u_{jkl}^{ABC}$ are referred to as interaction terms. Model (1) is referred to as the saturated model. Certain restrictions are placed on the interaction terms in order to avoid over parameterization problems, see Christensen (1997).

Two models of interest related to Equation (1) are:

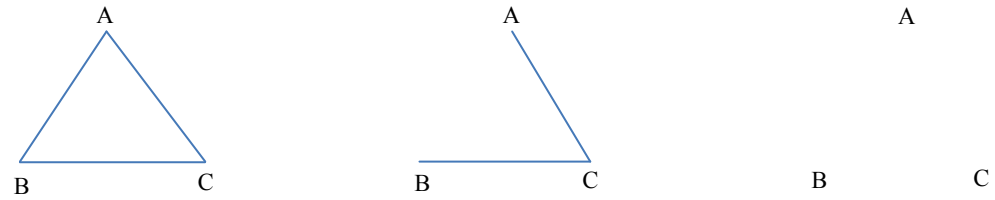
$$\log(m_{jkl}) = u + u_j^A + u_k^B + u_l^C + u_{jl}^{AC} + u_{kl}^{BC} \tag{2}$$

and

$$\log(m_{jkl}) = u + u_j^A + u_k^B + u_l^C \tag{3}$$

Model (2) specifies conditional independence, i.e., given C, then A and B are independent. Model (3) specifies the situation where A, B, and C are independent.

Graphs are useful to show the dependency structure among variables as illustrated below.



(a) Dependent (b) Conditional Independence (c) Independence

Figure 1 A Graphical Presentation of Dependency among Three Categorical Variables

The interested reader may refer to David Edwards (2000), S. L. Lauritzen (1996) or J. Whittaker (1990) for further exposition on Graphical Models.

Next, we consider variables measured on a continuous scale.

3. Continuous Variables

Continuous variables are labeled as $X,Y,Z\ldots$, or as $X_1,X_2,X_3\ldots$. The classical distribution for a finite set of continuous variables is the multivariate normal distribution, see Anderson (2003). The number of variables is denoted by q.

Data is given by a $n\times q$ Table of observations with n denoting the number of rows of the Table. The entry in the i[th] row and j[th] column is denoted by $x_{ij}, 1 \le i \le n,\ 1 \le j \le q$.

The density of a multivariate normal is

$$\begin{aligned} f(x;\mu,\Sigma) &= N(\mu,\Sigma) \\ &= \frac{1}{(2\pi)^{q/2}} \det(\Sigma)^{-1/2} \exp\{-\frac{1}{2}(x-\mu)' \Sigma^{-1}(x-\mu)\} \end{aligned} \tag{4}$$

where x and μ are $q\times 1$ vectors, and $\sum$ is a $q\times q$ matrix. The parameters of the distribution are μ, the mean vector, and $\sum$, the variance-covariance matrix.

In case of q = 3, we have

$$\mu = \begin{pmatrix} E(X) \\ E(Y) \\ E(Z) \end{pmatrix} \quad \textbf{and} \quad \Sigma = \begin{pmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,Z) \\ \mathrm{cov}(Y,X) & \mathrm{var}(Y) & \mathrm{cov}(Y,Z) \\ \mathrm{cov}(Z,X) & \mathrm{cov}(Z,Y) & \mathrm{var}(Z) \end{pmatrix} \tag{5}$$

If $\sum$ is a diagonal matrix then X, Y, and Z are independent. Y and Z are conditionally independent given X if the partial correlation of Y and Z given X denoted by $\rho_{Y,Z,X}$ is zero or equivalently if the Cov(Y, Z) = 0, see David Edwards (2000). Complete dependency among X,Y,Z is provided when all the elements of $\sum$ are non-zero.

In many real life situations, the multivariate normal family of distributions is too restrictive to define the joint distribution for a finite set of continuous variables. The following three procedures tend to mitigate this problem: (1) by considering a mixture of multivariate distribution, see McLachlan and Peel (2000); (2) using marginal distributions in conjunction with a specified copula to construct a multivariate joint distribution, refer to Nelsen (2006); and (3) use a multivariate version of Box–Cox transformation, as given in Johnson and Wichern (2007).

Next we consider variables of mixed types, i.e., some categorical and some continuous.

4. Models for Mixed Variables

Mixed variables problems involve the study of categorical as well as quantitative (continuous) variables. For example, consider the case where the variables are A, B, X, Y, Z with A and B as categorical (p = 2) variables, and X, Y, and Z as continuous (q = 3) variables.

Before we give an expression for the joint probability distribution of A, B, X, Y, Z, we need to introduce some necessary notations. The frequency Table associated with A, B will have #(A).#(B) distinct cells. For instance, if A can take values in the set {1,2} and B can take values in the set {1,2,3} then there are (2).(3) = 6 possible cell labels. A typical cell address is labeled as i. Furthermore; we shall designate a possible value of the triplet X, Y, Z by $w = (x,y,z)$.

The joint distribution for A, B, X, Y, Z is

$$\begin{aligned} f(i,w) &= f(i)\, f(w \mid i) \\ &= p_i\, N(\mu_i,\Sigma_i) \\ &= p_i\, \frac{1}{(2\pi)^{q/2}} \det(\Sigma_i)^{-1/2} \exp\{-\frac{1}{2}(w-\mu_i)' \Sigma_i^{-1}(w-\mu_i)\} \end{aligned} \tag{6}$$

where p_i is the probability for the cell i, and $N(\mu_i, \sum_i)$ is a multivariate normal of dimension q. In the example above q = 3. Note that for each i, there is a corresponding multivariate normal distribution $N(\mu_i, \sum_i)$ whose parameters depend upon i.

Here, the statistical issues of interest are addressed by reference to a simple example. Let us consider the case when we have one categorical variable A, and one continuous variable X, i.e., with p = 1, and q = 1. We shall write AX to designate this pair.

The data used for the statistical analysis of AX appears in Edwards (2000, Table 4.2, p. 70). In this instance, A represents the "type of diet", with four different diet types; and X denotes realization of "coagulation time (seconds) for blood drawn" from 24 animals randomly allocated to different diets. The data is reproduced in the Appendix below.

The joint density of AX is given as

$$f(i,x) = p_i \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\{-\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2\} \tag{7}$$

with i = 1,2,3,4.

There are two statistical problems of interest for this example. The first problem is concerned with the estimation of parameters of interest namely $p_i's, \mu_i's, and\ \sigma_i^2$. The second problem relates to study of the nature of dependency between A and X.

The estimation of parameters is based on the method of Maximum Likelihood, ML. ML estimation is based on minimizing the negative of the log of likelihood function. The ML estimation requires solving a system of nonlinear equations whose solution is implemented by an appropriate algorithm. In the Appendix, we have provided the necessary code to accomplish this task using an ADMB program.

The dependency structure of A and X can be examined by performing a number of model comparisons. The model comparison is based on Likelihood Ratio Test, LRT.

If M_r (reduced) is a model nested within M_f (saturated) model, then the large sample Likelihood Ratio test statistics is

$$2\{[-\log(Likelihood_{M_r})] - [-\log(Likelihood_{M_f})]\} \tag{8}$$

The asymptotic distribution of LRT is a Chi-square distribution with degrees of freedom equal to difference in the number of parameters in the two competing models.

The results for our model comparison are given in Table 3 below.

Table 3 Model Comparison

Case	Negative log of Likelihood	Comparison of Cases	Value of LRT statistic	P-value
(1) Same $\mu_j = \mu$ &same $\sigma_j = \sigma$	98.4567	1 vs. 4	29.1556	0.000057
(2) Different $\mu_j's$ & *same* σ_j	85.1313	2 vs. 4	2.5048	0.474400
(3) Same $\mu_j = \mu$ & different $\sigma_j's$	97.2028	3 vs. 4	26.6478	0.000007
(4) Different $\mu_j's$ & $\sigma_j's$	83.8789		NA*	NA*

Note: * Not Applicable.

Case 4, in Table 3 above, presents the saturated ("largest") model in our example. By contrast, Case 1 presents the "smallest" model in our example.

If Case 1 is valid then we have the same normal distribution for each value of *i*, *i* = 1,2,3 or 4. It implies that

A and X are statistically independent in this instance. The LRT used for comparing Case 1 to Case 4 has a value of 29.1556 with a p-value of 0.000057 which is extremely small suggesting the data does not support the hypothesis that A and X are independent.

Comparing Case 2 with Case 4, the LR test statistics is 2.5048 with a large p-value of 0.474400. Based on the 5% significance level, then we cannot reject the hypothesis that $\sigma_j's$ differ. Edwards (2000) refers to this situation as "homogeneity of variance", analogous to analysis of variance situation.

Finally, comparing Case 3 to Case 4, the LR test statistic has a very small p-value (0.000007) which rules out the hypothesis that the same μ can be utilized in for all diet types.

Table 3 above is helpful for studying the dependency structure between categorical and continuous variables in mixed variables settings.

5. Conclusions

This paper discussed briefly models for mixed variables. In particular, suitable joint distribution of mixed variables is provided by reference to a specific example. Procedures based on model comparison are utilized to study the dependency structure pertaining to a categorical and a continuous variable by reference to a simple example. Estimation of parameters and computation of Likelihood function was addressed by providing the necessary algorithm using ADMB code.

References:

ADMB project, available online at: http://www.admb-project.org.
Agresti Alan (2002*). Categorical Data Analysis* (2nd ed.), Wiley.
Anderson T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.), Wiley.
Christensen Ronald (1997). *Log-Linear Models and Logistic Regression* (2nd ed.), Springer.
Edwards David (2000). *Introduction to Graphical Modeling*, Springer.
Fienberg Stephen E. (1980). *The Analysis of Cross-Classified Categorical Data* (2nd ed.), Cambridge, MA: MIT Press.
Johnson Richard A. and Wichern Dean W. ((2007). *Applied Multivariate Statistical Analysis* (6th ed.), Pearson-Prentice Hall.
Lauritzen S. L. (1996). *Graphical Models*, Clarendon Press, Oxford.
McLachlan G. and Peel D. (2000). *Finite Mixture Models*, Wiley
Nelsen Roger B. (2006). *An Introduction to Copulas* (2nd ed.), Springer.
Whittaker J. (1990). *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons.

Appendix: ADMB Program for MLE of Parameters (Case 4 of Table 3)

Part 1 – ADMB Program Code

```
DATA_SECTION
  init_int nobs
  init_vector a(1,nobs)
  init_vector x(1,nobs)
PARAMETER_SECTION
  init_bounded_number p1(0,.3); init_bounded_number p2(0,.3); init_bounded_number p3(0,.3)
  number p4
  init_number mu1; init_number mu2; init_number mu3; init_number mu4
  init_number logs1; init_number logs2; init_number logs3; init_number logs4
  sdreport_number s1; sdreport_number s2; sdreport_number s3; sdreport_number s4
  number l1; number l2; number l3; number l4
  objective_function_value f
PROCEDURE_SECTION
```

```
   int i; const double pi=3.14159265359;
   p4 = 1-p1-p2-p3;
   s1 = exp(logs1); s2=exp(logs2);   s3 = exp(logs3); s4 = exp(logs4);
   l1=0.0;   l2=0.0;   l3=0.0;   l4=0.0;
   for (i = 1; i <= nobs; i++){if(a(i)==1) l1=l1+log(p1)-0.5*log(2*pi*s1*s1)-(0.5/square(s1))*square(x(i)-mu1);
if(a(i)==2) l2 = l2+log(p2)-0.5*log(2*pi*s2*s2)-(0.5/square(s2))*square(x(i)-mu2);
if(a(i)==3) l3 = l3+log(p3)-0.5*log(2*pi*s3*s3)-(0.5/square(s3))*square(x(i)-mu3);
if(a(i)==4) l4=l4+log(p4)-0.5*log(2*pi*s4*s4)-(0.5/square(s4))*square(x(i)-mu4);}
   f = -(l1+l2+l3+l4);
```

Part 2-Data

```
24
1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 4
62 60 63 59 63 67 71 64 65 66 68 66 71 67 68 68 56 62 60 61 63 64 63 59
```

Part 3-Parameter estimates & value of the objective function (Formatted ADMB Output)

```
# Number of parameters = 11   Objective function value = 83.8789   Maximum gradient component = 2.41911e-005
# p1:     0.166666740293
# p2:     0.249999794993
# p3:     0.249999832877
# mu1: 60.9999996189
# mu2: 66.0000003949
# mu3: 67.9999999312
# mu4: 60.9999994702
# logs1: 0.458146186021
# logs2: 0.948560431158
# logs3: 0.423646914273
# logs4: 0.895879132220
```