Workshop on Bayesian Modeling Using WinBUGS

Sessions 8–9 (B): Dummy variables and ANOVA models

Workshop web-page: http://www.math.ntua.gr/~fouskakis/Workshop/home.html

Workshop on Bayesian Modeling Using WinBUGS

Sessions 8–9 (B): Dummy variables and ANOVA models

Ioannis Ntzoufras

Department of Statistics,
Athens University of Economics and Business,
Athens, Greece;

e-mail: ntzoufras@aueb.gr.

Workshop web-page: http://www.math.ntua.gr/~fouskakis/Workshop/home.html

This presentation is based on Chapter 6 of Ntzoufras (2009): *Bayesian Modeling Using WinBUGS*, Wiley.

Synopsis

Parametrizations and Dummy Variables

1. ANOVA models using dummy variables
   - An One-way example in WinBUGS
   - ANOVA models using dummy variables - A 3-way example in WinBUGS
2. Analysis of covariance models
   - One quantitative variable and one qualitative variable
   - The parallel lines model
   - The separate lines model
   - A bioassay example in WinBUGS
   - Slope ratio analysis: Models with common intercept and different slope
3. Further modeling issues
   - Extending the simple ANCOVA model
   - Using binary indicators to specify models in multiple regression
   - Selection of variables using the deviance information criterion (DIC)
   - A stepwise method for DIC based variable selection in WinBUGS
4. Closing remarks

Introduction

Here we focus on

- The incorporation of categorical covariates in the usual regression models
- The use of dummy variables for the CR and STZ parametrizations
- ANOVA models using dummy variable

Dummy variables

- Categorical variables can be incorporated in regression models via the use of dummy variables.
- These dummy variables identify which parameters must be added (and how) to the linear predictor.

Dummy variables for CR parametrization

Consider the simple one-way ANOVA with a categorical factor A and L_A categories (CR parametrization):

$$\mu_i = \mu_0 \quad \text{when} \quad a_i = 1$$
$$\mu_i = \mu_0 + \alpha_{a_i} \quad \text{when} \quad a_i > 1.$$

AIM: Express μ_i as a linear combination of covariates:

$$\mu_i = \mu_0 + \alpha_2 D^A_{i2} + \alpha_3 D^A_{i3} + \cdots + +\alpha_{L_A} D^A_{iL_A},$$

$D^A_{i\ell}$; $\ell = 1, 2, \ldots, L_A$ are dummy variables:

$$D^A_{i\ell} = 1 \;\text{ if }\; a_i = \ell \;\text{ and }\; D^A_{i\ell} = 0 \;\text{ otherwise.} \tag{1}$$

- $(\mu_0, \alpha_2, \ldots, \alpha_{L_A})$ play the same role as $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ in regression
- Number of dummies: $L - 1$, with L =number of levels

Dummy variables for STZ parametrization

- For the STZ parametrization:

$$\mu_i = \mu_0 - \alpha_2 - \alpha_3 \cdots - \alpha_{L_A} \quad \text{when} \quad a_i = 1$$
$$\mu_i = \mu_0 + \alpha_{a_i} \quad \text{when} \quad a_i > 1.$$

- The dummy variables here are slightly more complicated since

$$\begin{aligned} D^{A,\text{stz}}_{i\ell} &= \;\;\,1 \quad \text{if} \quad a_i = \ell \\ D^{A,\text{stz}}_{i\ell} &= -1 \quad \text{if} \quad a_i = 1 \\ D^{A,\text{stz}}_{i\ell} &= \;\;\,0 \quad \text{otherwise.} \end{aligned} \tag{2}$$

- Here, first category = function of the remaining of the parameters.
- If we wish to omit from the parameters a different category, then the dummy variables must be modified accordingly.

Dummy variables - additional details

- STZ dummy variables can be expressed as the difference of the CR dummy variables:
$$D_{i\ell}^{A,\mathrm{stz}} = D_{i\ell}^{A} - D_{i1}^{A} \ . \qquad (3)$$
- In CR parametrization: The reference level is specified by the dummy that we remove from the linear predictor.
- The use of dummy variables simplifies the model specification:
  - The structure is the same (linear) regardless of the type of parametrization or data we use.
  - Only actually used parameters are defined in the model.
  - Omitted parameters (as in the STZ parametrization) can be directly monitored using a simple deterministic/logical node in WinBUGS model code.

Interaction terms and dummy variables

Interaction terms are defined by the products of the corresponding dummy variables.

For example, the product $D_{ia}^{A} D_{ib}^{B}$ will provide the parameter $\alpha\beta_{ab}$ of the interaction between factors A and B for the corner constraints. Hence we can write
$$D_{iab}^{AB} = D_{ia}^{A} D_{ib}^{B}, \qquad (4)$$
where D_{iab}^{AB} is the dummy variable for the interaction between a and b levels of factors A and B for i subject.

- This multiplicative property is also true for higher interaction terms
- It is convenient in terms of WinBUGS programming since the definition of the individual constraints is avoided.

Specification of Dummy Variables in WinBUGS

Dummies for CR parametrization

(1st level as a reference category)

```
for (i in 1:n){
    D.A2[i] <- equals(a[i],2)
    ....
    D.ALA[i] <- equals(a[i],LA)
}
```

where `LA` is the number of levels of factor A, L_A.

Alternative approach using a matrix of dimension $n \times L_A$:

```
for (i in 1:n){
    for (l in 1:LA){
         D.A[i,l] <- equals(a[i],l)
}}
```


Dummies for STZ parametrization

(omitting the 1st level)

Using this syntax, we set $D.A_{i\ell} = 1$ if $a_i = \ell$ and zero otherwise following

```
for (i in 1:n){
    for (l in 1:LA){
         DSTZ.A[i,k] <- equals(a[i],l) - equals( a[i], 1 )
    }}
```

produces the corresponding dummies for the STZ parametrization.

Alternatively: Define first the CR dummies and then set STZ dummies as their difference:

```
         DSTZ.A[i,l] <- D.A[i,l] - D.A[i,1]
```

Dummy variables - additional comments

- Alternatively, dummy variables can be defined in another statistical package and then imported in WinBUGS model code as ready-to-use data.
- The $n \times p$ matrix $\boldsymbol{X}$ (of dimension $n \times p$) with columns
  - the constant term,
  - the dummy variables of the factors, and
  - the corresponding interactions involved in the model

  is called a *design matrix*.
- When mixed (dummy and continuous) types of variables are included in the linear predictor, it is called a *data matrix* as in the regression model.
- We can use matrix $\boldsymbol{X}$ to specify our model as described in the usual regression model.
- Such a strategy $\Rightarrow$ considerably simplifies the model specification.

- The design matrix can be easily constructed either within WinBUGS model code (following similar approach as for the construction of dummy variables) or outside WinBUGS and then by importing the design matrix in the data of the WinBUGS code; for a detailed example, see Section 6.1.
- Since matrix $\boldsymbol{X}$ can be defined regardless of the the type of variables, we can use all prior distributions described in regression models.
- The algebra related to dummy variables is intriguing since, in cases of balanced data, all design matrices can be calculated using Krönecker products.
- STZ parametrization has interesting properties resulting in independent posterior distributions of parameters related to different model terms (main effects and interaction parameters); see Ntzoufras (1999) for a brief discussion.

6.1 Analysis of variance models using dummy variables

Example 5.2 (continued): Evaluation of tutors.

Table of Dummies

Tutor	D_1	D_2	D_3	D_4	D_2^{STZ}	D_3^{STZ}	D_4^{STZ}
1	1	0	0	0	-1	-1	-1
2	0	1	0	0	1	0	0
3	0	0	1	0	0	1	0
4	0	0	0	1	0	0	1

WinBUGS Code - CR

```
model{
    # Dummies
    for (i in 1:n){ for (j in 1:LA) {
        # CR PARAMETRIZATION
        # D[i,j] <- equals( class[i], j )
        # STZ PARAMETRIZATION
          D[i,j] <- equals( class[i], j ) - equals(class[i], 1 )
    }}
    #
    # model's likelihood
    for (i in 1:n){
        mu[i] <- m + inprod( alpha[2:LA], D[i,2:LA] )
        grade[i] ~ dnorm( mu[i], tau )
    }
    #### stz constraints
     alpha[1] <-  -sum(alpha[2:LA])
    #### CR Constraints
```

```
    # alpha[1] <- 0.0

    # priors
    m~dnorm( 0.0, 1.0E-04)
    for (i in 2:LA){ alpha[i]~dnorm(0.0, 1.0E-04)}
    tau ~dgamma( 0.01, 0.01)
    s <- sqrt(1/tau) # precision
}

INITS
list( m=1.0, alpha=c(NA, 0,0,0), tau=1.0 )

DATA (LIST)
list( n=25,  LA=4,
      grade=c(84, 58, 100, 51, 28, 89, 97, 50,  76, 83, 45, 42, 83,
              64, 47,  83, 81, 83, 34, 61, 77, 69,  94, 80, 55, 79),
      class=c(1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2,
              3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4) )
```


Results using Dummies for STZ

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha[1]	-0.595	7.933	0.0726	-16.32	-0.587	14.97	1001	6000
alpha[2]	-1.319	7.429	0.1202	-16.1	-1.353	13.43	1001	6000
alpha[3]	-4.082	7.508	0.1035	-18.6	-4.042	10.82	1001	6000
alpha[4]	5.996	8.398	0.1284	-10.6	5.962	22.76	1001	6000
m	68.86	4.491	0.047	59.81	68.84	77.99	1001	6000

WinBUGS Code - STZ

```
model{
    # Dummies
    for (i in 1:n){ for (j in 1:LA) {
        # CR PARAMETRIZATION
        # D[i,j] <- equals( class[i], j )  ← remove from comment
        # STZ PARAMETRIZATION                change this to comment
          D[i,j] <- equals( class[i], j ) - equals(class[i], 1 )  ↩
    }}
    for (i in 1:n){
        mu[i] <- m + inprod( alpha[2:LA], D[i,2:LA] )
        grade[i] ~ dnorm( mu[i], tau )
    }
    #### stz constraints
     alpha[1] <-  -sum(alpha[2:LA])  ← remove from comment
    #### CR Constraints
    # alpha[1] <- 0.0                 ← change this to comment
    ........................................................
```


WinBUGS Code - STZ

```
model{
    # Dummies
    for (i in 1:n){ for (j in 1:LA) {
        # CR PARAMETRIZATION
          D[i,j] <- equals( class[i], j )
        # STZ PARAMETRIZATION
        # D[i,j] <- equals( class[i], j ) - equals(class[i], 1 )
    }}
    for (i in 1:n){
        mu[i] <- m + inprod( alpha[2:LA], D[i,2:LA] )
        grade[i] ~ dnorm( mu[i], tau )
    }
    #### stz constraints
    # alpha[1] <-  -sum(alpha[2:LA])
    #### CR Constraints
     alpha[1] <- 0.0
    ........................................................
```

Results using Dummies for CR

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha[2]	-0.2027	12.38	0.4066	-24.71	-0.1435	24.74	1001	5000
alpha[3]	-3.132	12.19	0.4014	-27.24	-3.094	21.22	1001	5000
alpha[4]	6.867	13.52	0.4154	-19.81	6.824	34.08	1001	5000
m	67.95	8.931	0.33	50.91	68.1	85.72	1001	5000
s	22.19	3.617	0.05842	16.49	21.74	30.68	1001	5000

Defining data in Dummies

- Remove code for dummies
- Use the following data

```
DATA (LIST)
list( n=25,  LA=4)
CR DATA (rect)
grade[]   D[,1]   D[,2]   D[,3]   D[,4]
```

For CR:

```
84   1 0 0 0
...........
89   1 0 0 0
97   0 1 0 0
...........
83   0 1 0 0
64   0 0 1 0
...........
61   0 0 1 0
77   0 0 0 1
...........
79   0 0 0 1
END
```

For STZ:

```
84   0 -1   -1   -1
...............
89   0 -1   -1   -1
97   0  1    0    0
...............
83   0  1    0    0
64   0  0    1    0
...............
61   0    0  1    0
77   0    0  0    1
...............
79   0  0    0    1
END
```

Levels

1
...
1
2
...
2
3
...
3
4
...
4

- Use the following code for `alpha[1]`
  - For CR:

```
alpha[1] <-  -sum(alpha[2:LA])
```

  - For STZ:

```
alpha[1] <- 0.0
```


References

Ntzoufras, I. (1999), “Discussion on Bayesian model averaging and model search strategies”, in J. Bernardo, J. Berger, A. Dawid, and A. Smith, eds., *Bayesian Statistics*, Vol. 6, Oxford University Press, pp. 178–179.

Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*, Wiley Series in Computational Statistics, Hoboken, NJ.