Full Dummy Coding for Nominal Variables in Bayesian Regression

Kevin S. Van Horn

June 30, 2015

Abstract

Full dummy coding is avoided in frequentist regression analyses because it leads to non-identifiability. It is not recommended for Bayesian analyses for a different reason: full dummy coding is computationally problematic. The group indicators for a hierarchical model are a special case that avoids these computational problems.

Consider a regression analysis with a vector v of input variables—which may include both continuous and discrete variables—that gets transformed into a vector x of predictor variables. For example, if v_1 is a continuous input variable and v_2 is a binary input variable having allowed values `yes` and `no`, then x could be a length-4 vector defined by

$$\begin{aligned} x_1 &= 1 \\ x_2 &= v_1 \\ x_3 &= v_1^2 \\ x_4 &= \begin{cases} +1 & \text{if } v_2 = \texttt{yes} \\ -1 & \text{if } v_2 = \texttt{no} \end{cases} \end{aligned}$$

and the dependent variable would depend on v through the dot product[1]

$$\beta' x = \sum_j \beta_j x_j$$

where the β is the vector of regression coefficients (β_1 is the intercept in this case).

[1] We write β' for the transpose of β; that is, the row vector corresponding to the column vector β.

In this note we are concerned with the encoding of nominal variables: v_j is a nominal variable with K possible levels, and

$$x = (1, \ldots, \tau_K(v_j), \ldots)'$$

where τ_K gives the encoding of a K-level nominal input variable. One possibility for τ_K is *full dummy coding*:[2]

$$\begin{aligned} \tau_K(v_j) &= (w_1, \ldots, w_K) \\ w_k &= (v_j = k). \end{aligned}$$

In frequentist analyses this full dummy coding is not used if there is a separate intercept, because it yields a non-identifiable model: subtracting a constant amount from the regression coefficient for each w_k and adding it to the intercept leaves the likelihood unchanged. The same problem occurs if there are multiple nominal variables and each uses full dummy coding. Instead, a leave-one-out dummy coding is used (w_K is dropped, so that level K is encoded as all zeroes) or an effects coding is used (w_K is dropped, but $w_k = -1$ for all $1 \leq k \leq K-1$ when $v_j = K$).

In principle, non-identifiability is not a problem for a Bayesian analysis: we are interested in obtaining a posterior distribution for the regression coefficients, not a point estimate, and as long as we use a proper prior distribution we are guaranteed a proper posterior distribution. Thus we could, if desired, use a full dummy coding.

In practice, things are less straightforward. Use of full dummy coding in a Bayesian analysis can be problematic because it yields a highly correlated posterior distribution that poses *computational* difficulties.

To see this, consider a simple problem: a linear regression with only a single input variable v, which is a nominal variable with K levels. We'll modify our notation to accomodate multiple data points: v_i is now the value of this input variable for the i-th data point, and gets transformed into the row vector X_i (so that X is a matrix with one row for each observation.) y_i is the value of the dependent variable for the i-th data point. We have

$$y_i \sim \text{Normal}(X_i\beta,\, \sigma_y), \quad 1 \leq i \leq n$$

together with independent normal priors on the regression coefficients

$$\beta_j \sim \text{Normal}(0, \sigma_j).$$

[2] If φ is a logical expression, then (φ) is 1 when φ is true, and 0 when φ is false.

Define Σ_0 to be the diagonal prior covariance matrix and Λ_0 its inverse:

$$\begin{aligned}\Sigma_{0jk} &= \begin{cases}\sigma_j^2 & \text{if } j = k\\ 0 & \text{otherwise.}\end{cases}\\ \Lambda_0 &= \Sigma_0^{-1}.\end{aligned}$$

The predictor matrix X, using full dummy coding and 0-based indexing for the columns, is

$$\begin{aligned}X_{i0} &= 1 \quad (\text{so that } \beta_0 \text{ is the intercept})\\ X_{ij} &= (v_i = j) \quad \text{when } j \neq 1.\end{aligned}$$

The posterior distribution for β is then a multivariate normal distribution with covariance matrix Σ, where

$$\begin{aligned}\Sigma &= \Lambda^{-1}\\ \Lambda &= \sigma_{\mathrm{y}}^{-2} X'X + \Lambda_0.\end{aligned}$$

Looking at individual elements of Λ we have

$$\Lambda_{jk} = \Lambda_{0jk} + \sigma_{\mathrm{y}}^{-2} \sum_{i=1}^{n} X_{ij} X_{ik}.$$

If p_j is the proportion of rows X_i for which $X_{ij} = 1$, then

$$\begin{aligned}\Lambda_{j0} &= (j = 1)\sigma_1^{-2} + \sigma_{\mathrm{y}}^{-2} \sum_{i=1}^{n} X_{ij}\\ &= \begin{cases} n\sigma_{\mathrm{y}}^{-2} + \sigma_1^{-2} & \text{if } j = 0\\ np_j\sigma_{\mathrm{y}}^{-2} & \text{if } j \neq 0\end{cases}\end{aligned}$$

and for $j, k \neq 0$,

$$\begin{aligned}\Lambda_{jk} &= (j = k)\sigma_j^{-2} + \sigma_{\mathrm{y}}^{-2} \sum_{i=1}^{n} X_{ij} X_{ik}\\ &= (j = k)\left(np_j\sigma_{\mathrm{y}}^{-2} + \sigma_j^{-2}\right).\end{aligned}$$

That is,

$$\Lambda = n\sigma_{\mathrm{y}}^{-2} \begin{pmatrix} 1 & p_1 & p_2 & \cdots & p_K \\ p_1 & p_1 & 0 & \cdots & 0 \\ p_2 & 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_K & 0 & 0 & \cdots & p_K \end{pmatrix} + \begin{pmatrix} \sigma_0^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_1^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_K^{-2} \end{pmatrix},$$

and the left summand is a singular matrix: $(1, -1, \cdots, -1)'$ is an eigenvector with eigenvalue of 0. Defining

$$r_j = \frac{\sigma_{\mathrm{y}}^2}{n\sigma_j^2}$$

we see that if $r_j \ll 1$ for all j then Λ is nearly singular.

A nearly singular posterior precision matrix Λ is problematic even in this simple scenario, as it leads to numerical instabilities when inverting the matrix. It also implies very high posterior correlations among the regression coefficients. In generalized linear models where we have to use Markov chain Monte Carlo methods to explore the posterior distribution—say, logistic regression or Poisson regression—this leads to very slow mixing of the Markov chain; in the case of Hamiltonian Monte Carlo it can lead to very small step sizes. In either case the result is very slow estimation.

As an example, consider a data set with $n = 100$ observations and a weak prior, $\sigma_j = 10^3\sigma_{\mathrm{y}}$. Then $r_j = 10^{-8}$, and for $K = 3$ we find that the posterior correlation between β_0 and β_j ($j \neq 0$) is

$$\rho_{0j} = -0.99999994$$

and the posterior correlation between β_j and β_k, $j, k \neq 0$, $j \neq k$, is

$$\rho_{jk} = 0.9999999.$$

How large does r_j need to be to avoid extreme correlations? Assuming $\sigma_j = \sigma_k$ for all $j, k \neq 1$, and hence $r_j = r_k$ for all $j, k \neq 0$, the following table gives the minimum value that r_j (for $j \neq 0$) may have and still yield posterior correlations that never exceed 0.9 in absolute value. These figures assume $r_0 = 10^{-8}$ (they are not sensitive to the specific value of r_0 used.)

K	min r_j
3	2.6×10^{-2}
4	1.5×10^{-2}
5	9.4×10^{-3}
6	6.5×10^{-3}
8	3.7×10^{-3}
10	2.3×10^{-3}
13	1.4×10^{-3}
16	9.2×10^{-4}
19	6.5×10^{-4}

Notice that as the number of levels K increases, smaller values for r_j become acceptable.

Hierarchical models can be thought of as regression models involving a nominal variable with many distinct levels. The simplest case is a varying-intercept model:

$$\begin{aligned} y_i &\sim \text{Normal}\,(\upsilon_i, \sigma_y) \\ \upsilon_i &= \alpha + a_{k[i]} + \beta x_i \end{aligned}$$

where $k[i]$ is the group to which observation i belongs, and we have the hierarchical prior

$$\begin{aligned} a_k &\sim \text{Normal}\,(0, \sigma_a) \\ \sigma_a &\sim \text{some diffuse prior.} \end{aligned}$$

The equation for υ_i is equivalent to

$$\begin{aligned} \upsilon_i &= \sum_{k=1}^{K} a_k w_{ik} + \beta x_i \\ w_i &= \text{full dummy coding of level } k[i]. \end{aligned}$$

So why don't we see a problem with high posterior correlations among the coefficients a_k in this case? We have two mitigating factors in play:

- K is large. As previously noted, this allows smaller r_k values without getting extreme posterior correlations.
- σ_a is relatively small, yielding larger values for r_k. The whole point of a hierarchical model is that we expect to find the posterior distribution for σ_a to be concentrated on values much smaller than what we would assign for a weakly informative prior on the coefficients a_k.

In summary, we have the following guidelines for encoding nominal variables in a Bayesian regression analysis:

- Do not use full dummy coding, as it will generally lead to severe computational difficulties due to extreme posterior correlations in the regression coefficients.
- Group indicators for hierarchical models are a special case that avoids these difficulties.