ARTICLE IN PRESS

Journal of Choice Modelling xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm

Dummy coding vs effects coding for categorical variables: Clarifications and extensions

Andrew Daly, Thijs Dekker, Stephane Hess*

Institute for Transport Studies & Choice Modelling Centre, University of Leeds, United Kingdom

ARTICLE INFO

Keywords:
Dummy coding
Effects coding
Discrete choice
Categorical variables

ABSTRACT

This note revisits the issue of the specification of categorical variables in choice models, in the context of ongoing discussions that one particular normalisation, namely effects coding, is *superior* to another, namely dummy coding. For an overview of the issue, the reader is referred to Hensher et al. (2015, see pp. 60–69) or Bech and Gyrd-Hansen (2005). We highlight the theoretical equivalence between the dummy and effects coding and show how parameter values from a model based on one normalisation can be transformed (after estimation) to those from a model with a different normalisation. We also highlight issues with the interpretation of effects coding, and put forward a more well-defined version of effects coding.

1. Introduction

Choice models describe the utility of an alternative as a function of the attributes of that alternative. Those attributes might be continuous or categorical, where the simplest example of the latter is a binary present/absent attribute. In the case of continuous attributes, the associated coefficient (say β) measures the marginal utility of changes in the attribute (say x), and the analyst simply needs to make a decision on whether that marginal utility is independent of the base level of the attribute (i.e. linear) or not (i.e. non-linear). For the latter, we would simply replace βx in the utility function with $\beta f(x)$, where $f(x)$ is a non-linear transformation of x.

Just as for continuous attributes, an analyst needs to make a decision on how categorical variables are treated in the computation of an alternative's utility. For continuous variables, it is clear that different assumptions in terms of functional form (linear vs non-linear, and the specific degree of non-linearity) will lead to different model fit and behavioural implications. However, in the case where x is a categorical variable with a finite number of different levels, neither a linear nor a function-based non-linear treatment is likely to be appropriate. With x taking the values $1, \ldots, K$ for the K different levels it can assume, a linear treatment would imply that the difference in utility between successive levels are equal. The fundamental differences in meaning of different levels of categorical variables may also make it difficult for an analyst to find an appropriate functional form for $f(x)$ in a non-linear specification.

For categorical variables, model estimation captures the impact on utility of moving from one level of the attribute to a different level. Because of linear dependency, it is then impossible to estimate a separate parameter for each level. Different approaches have been put forward in the literature to address this issue, in particular dummy coding and effects coding, although many other approaches would be possible. The purpose of this note is to address ongoing discussions in the choice modelling literature which claim that some of the ways of dealing with such variables (in particular effects coding) are *superior* to others. For an overview of the issue, the reader is referred to Hensher et al., (2015, see pp. 60–69) or Bech and Gyrd-Hansen (2005).

* Corresponding author.
E-mail addresses: andrew@alogit.com (A. Daly), T.Dekker@leeds.ac.uk (T. Dekker), s.hess@its.leeds.ac.uk (S. Hess).

http://dx.doi.org/10.1016/j.jocm.2016.09.005

Available online xxxx

2. Equivalence and interchangeability of dummy and effects coding of categorical variables

We start by noting that the most widely found example of a categorical variable is that of an alternative-specific constants (ASC). These ASCs capture the mean of the error term for the utility of an alternative in a choice model. In the case of J different alternatives, we can easily specify J different binary variables for alternative j, say $z_{j,k}$, with $k = 1,\ldots,J$, where, for a given alternative j, $z_{j,k}$=1, if $j = k$, and 0 otherwise. It is important to stress that this applies both in the case of labelled alternatives, where e.g. $z_{j,k}$ may refer to a given mobile phone model, or unlabelled alternatives, where e.g. $z_{j,k}$ may refer to whether an alternative is presented on the left hand side in a survey. Not including ASCs is equivalent to an assumption that the means of the error terms, including unmeasured variables, are identical across alternatives. With this notation, the contribution of the ASC component to the utility of alternative j would now be given by $\sum_{k=1}^{J}\delta_k z_{j,k}$, where, for any given alternative j, only one element in $z_{j,k}$, with $k = 1,\ldots,J$, will be equal to 1 (and all others will be 0).

It is then widely known that a restriction has to be imposed to make the model identifiable, i.e. it is impossible to estimate a value δ_k, $\forall\ k$, where $k = 1, \ldots J$. This is a result of the fact that probabilities in choice models are defined in terms of utility differences across the J alternatives in the consumer's choice set, and an infinite number of combinations of values for the ASCs would lead to the same differences. A constraint needs to be applied to avoid linear dependency and thereby make the model identifiable. A typical approach in this context is to set the ASC for one alternative (e.g. the last) to zero, e.g. δ_J=0, and estimate the remaining ones, i.e. δ_j, $\forall\ j < J$. As we will see later on, this is in effect using dummy coding for the ASCs, with the value of one ASC set to 0 and all others being estimated.

While the use of dummy coding for ASCs is commonplace, substantial discussion has gone into the appropriate specification to use for other categorical variables, which are our main focus. Let us consider, for example, the case of a categorical variable for alternative j, say x_j, where we assume that this variable has K different levels (i.e. possible values).[1] In the absence of an acceptable assumption about some continuous relationship between the different levels of x_j and the utility of alternative j, we are then in a situation where we want to estimate independent different utility components for its different levels. This leads to the need for recoding and/or, as mentioned above, for some constraints to be imposed on the specification. The key distinction made in the literature has been between dummy coding and effects coding. We will now illustrate how these can be arrived at through different ways of recoding the categorical variable. The rationale behind recoding is that we create a number of separate variables for the separate levels of each categorical variable, where different parameters are associated with the resulting variables.

Table 1 presents the way in which a four-level categorical variable would typically be recoded into dummy and effects coding (see also Hensher et al. (2015, see pp. 60–69) or Bech and Gyrd-Hansen (2005)). With either approach, we recode the categorical variable x into four new variables, which each have an associated parameter, i.e. $\beta_1..\beta_4$ for the dummy-coded variables or, in the alternative model specification, $\gamma_1..\gamma_4$ for the effects-coded variables.

When adopting either of these coding conventions, we can without loss of generality normalise the parameter of the fourth (last) level to zero for identification purposes.[2] We thus estimate three parameters with either approach. The difference arises in the final level of x. For the first *K-1* levels, one recoded variable is equal to 1 (e.g. with the first level, $D_{1,x}$ for dummy coding and $E_{1,x}$ for effects coding), while all others are zero. However, while this is still the case for the final level of x also for dummy coding, with effects coding, we additionally set the values for the first K recoded variables to −1. The utility contribution $\omega_{K,D}$ associated with level K under dummy coding is then zero since β_K=0, while, under effects coding $\omega_{K,E}=-\sum_{k<K}\gamma_k$, i.e. the negative sum of *all* the estimated effects coded parameters (remembering that γ_K=0). The sum of utility contributions across all K levels is thus zero with effects coding and this is the objective of effects coding.

The first point that we wish to make in this paper is that while the approach described above is the *'standard'* way of understanding effects coding in much of the literature, the recoding itself is not required. Indeed, a completely equivalent approach would be to retain the dummy coded variables but, instead of setting β_K=0, we could use a normalisation that $\beta_K=-\sum_{k<K}\beta_k$. This would then give us that $\beta_k=\gamma_k \forall\ k < K$, and $\omega_{k,D}=\omega_{k,E}$, $\forall\ k$. In short, the two alternative recoding approaches are identical, but rely on alternative normalisations. The recoding typically applied when analysts use effects coding is thus not required. Should we wish to have effects coded parameters, we can just recode the categorical variable into K binary 0–1 variables as in dummy coding, and impose the constraint $\beta_K=-\sum_{k<K}\beta_k$.

Moving away from the issue of recoding of variables, our next point relates to the equivalence between the two approaches. The above discussion highlights that the utility contribution of dummy and effects coding differs up to a constant, i.e. $\omega_D^k=\omega_E^k+\sum_{z<K}\gamma_z$. The latter term drops out when taking differences across levels leaving the utility difference between any two categories unchanged.

$$\omega_{k,D} - \omega_{l,D} = \omega_{k,E} - \omega_{l,E} \text{ for } \forall\ k,\ l \quad = 1..\ K$$

When level K is not contrasted then we get $\beta_k-\beta_l=\gamma_k-\gamma_l$ for $\forall\ k,\ l = 1.\ .K-1$. This directly follows from the utility contributions in Table 1. More important, we can use the contrast with level K to establish the correspondence between the two coding schemes.

$$\omega_D^k-\omega_D^K=\omega_E^k-\omega_E^K=\beta_k=\gamma_k+\sum_{l<K}\gamma_l$$

[1] Similar coding procedures would also be applicable when linearly dependent variables occur across alternatives, but for simplicity we focus here on the case of variables applying for a single alternative.

[2] In dummy coding correlations between the parameters can be minimised by selecting the base category to be the one most frequently observed. This step will additionally reduce the reported estimation errors of the parameters.