

# ULRR

## The choice of reference subclass in categorical regression models matters

Item Type	Meetings and Proceedings
Authors	Mackenzie, Gilbert;Peng, Defen
Citation	29th International Workshop on Statistical Modelling;
Download date	2026-03-09 00:33:55
Item License	<a href="https://creativecommons.org/licenses/by-nc-sa/1.0/">https://creativecommons.org/licenses/by-nc-sa/1.0/</a>
Link to Item	<a href="https://hdl.handle.net/10344/5272">https://hdl.handle.net/10344/5272</a>

# The choice of reference subclass in categorical regression models matters

Gilbert MacKenzie<sup>1</sup> and Defen Peng<sup>2</sup>

<sup>1</sup> ENSAI, Rennes, France

<sup>2</sup> University of British Columbia, Canada

Email: gilbert.mackenzie@ul.ie and defen.peng@ul.ie

**Keywords:** Categorical regression; Choice of reference subclass; Condition number; Generalized linear models; Multi-collinearity; Total variance.

## 1 Introduction

In the parametric regression models with categorical covariates, it is well known that many key quantities of interest are invariant to the choice of reference subclass. However, surprisingly, not all quantities are invariant and some choices may lead to models which have inferior properties when judged against particular criteria. We propose a set of secondary criteria upon which the choice of reference subclass may be based. This, secondary, set comprises: (a) precision of the estimates, (b) a measure of multi-collinearity and (c) subject matter considerations. The elements of this set are clearly inter-related. We explore the development and use of the proposed criteria in generalized linear models (GLMs) with categorical covariates. Our approach is based on analysis, simulation studies and a detailed analysis of a real data set. The results show clearly that it is possible to improve the characteristics of the model by selecting the reference subclass judiciously. This findings is based on the close relationship between the measure of precision of the estimates and the measure of multi-collinearity. So that it is natural to wish to evaluate any choice based on subject matter considerations in terms of the former two criteria.

Our approach is to develop a measure of the precision of the regression estimates,  $V_r$ , the total variance, and adopt a measure of the condition of  $V_r$ , namely  $K_r$ , and to consider the dependence between the pair  $(V_r, K_r)$  as we vary  $r$

## 2 General linear model with categorical covaiates

We consider some basic ideas and notation in the context of the general linear model with Gaussian responses  $Y$ . Later we generalize our findings

to the class of generalized linear models (GLMs). Classically, we assume  $E(Y) = X\beta$  where,  $Y$  is a continuous response variable,  $X$  is an  $n \times p$  design matrix,  $\beta$  is a  $p \times 1$  column vector of regression parameters. It follows immediately that  $\hat{\beta} = (X'X)^{-1}X'Y$  and that the covariance matrix  $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ .

Now let  $X$  be a single categorical variable with  $p = k + 1$  subclasses. In regression analysis, encoding the categorical variable,  $X$ , is conventional. Dummy variable coding which compares each subclass with the reference subclass, is the scheme most commonly adopted. Under this coding scheme, the columns of  $X$  comprises an intercept and  $k$  binary indicator variables. The vector of subclasses numbers is  $(n_1, n_2, \dots, n_p)$  and let the subscript  $r$ , denote the reference category which may be chosen freely from  $(1, \dots, p)$ .

## 2.1 Basic formulae

Then

$$X'X = \begin{pmatrix} n & n_1 & n_2 & \cdots & n_k \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ n_k & 0 & 0 & \cdots & n_k \end{pmatrix}, \quad (1)$$

whence we have

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i[1]} y_i \\ \sum_{i[2]} y_i \\ \vdots \\ \sum_{i[k]} y_i \end{pmatrix} \quad (2)$$

where  $i[j] = \{i \in j\text{th category} : x_{ji} = 1, j = 1, 2, \dots, r, \dots, p\}$ .

It follows that

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1} = \frac{\sigma^2}{n_r} \begin{pmatrix} +1 & -1 & -1 & \cdots & -1 \\ -1 & 1 + n_r/n_1 & +1 & \cdots & +1 \\ -1 & +1 & 1 + n_r/n_2 & \cdots & +1 \\ \vdots & \vdots & \vdots & & \vdots \\ -1 & +1 & +1 & \cdots & 1 + n_r/n_k \end{pmatrix} \quad (3)$$

for  $\hat{\beta} = (X'X)^{-1}X'Y$ , where  $n_r = n - \sum_{j=1}^k n_j$  is the number of observations allocated to the reference category. The matrix  $(X'X)^{-1}$  has a special structure which recurs in a more general form in GLMs (later) and again in the context of an interval-censored Exponential survival model (MacKenzie & Peng, 2013).

## 2.2 Two binary variables

The design matrix then has three columns  $(x_1, x_2, x_3)$  with corresponding definitions and subclass numbers  $(n_1, n_2, n_3)$  such that  $n_1 \neq n_2 \neq n_3$ . Again we first reparametrize  $X$  with three columns  $(x_0, x_1, x_2)$ , writing  $X \leftarrow (x_0, x_1, x_2)$ . Thus,  $X \leftarrow (x_0, x_1, x_2)$  implies that category 3 is the reference category since column  $x_3$  is omitted from the design matrix. Then

$$\hat{\beta}_{r=3} = \begin{pmatrix} \frac{1}{n_3} \sum_{i[3]} y_i \\ -\frac{1}{n_3} \sum_{i[3]} y_i + \frac{1}{n_1} \sum_{i[1]} y_i \\ -\frac{1}{n_3} \sum_{i[3]} y_i + \frac{1}{n_2} \sum_{i[2]} y_i \end{pmatrix}. \quad (4)$$

In a similar way we can easily find  $\hat{\beta}_{r=2}$  and  $\hat{\beta}_{r=1}$ . Consequently, we easily deduce that the estimates of the intercepts are all different and the pairs of regression coefficients  $(\hat{\beta}_{1,r}, \hat{\beta}_{2,r})$  for  $r = 3, \dots, 1$  are different for each value of  $r$ . Importantly, this latter finding differs from that found in the case of a single binary covariate.

For precision,

$$\begin{aligned} \text{diag } V(\hat{\beta}_{r=3}) &= \sigma^2 \left[ \frac{1}{n_3}, \left( \frac{1}{n_3} + \frac{1}{n_1} \right), \left( \frac{1}{n_3} + \frac{1}{n_2} \right) \right], \\ \text{diag } V(\hat{\beta}_{r=2}) &= \sigma^2 \left[ \frac{1}{n_2}, \left( \frac{1}{n_2} + \frac{1}{n_1} \right), \left( \frac{1}{n_2} + \frac{1}{n_3} \right) \right], \\ \text{diag } V(\hat{\beta}_{r=1}) &= \sigma^2 \left[ \frac{1}{n_1}, \left( \frac{1}{n_1} + \frac{1}{n_2} \right), \left( \frac{1}{n_1} + \frac{1}{n_3} \right) \right] \end{aligned} \quad (5)$$

and the diagonals are different for each  $r$ , unlike the single binary variable case. It should be noted, however, that the *pairing of values* pattern repeats in the variances.

For more than two binary covariates the patterns are similar, demonstrating that the single binary covariate is simply a rather special case. The key finding of this section is that, in general, the precision varies with  $r$  and from inspecting the denominators in (5) it is clear that the variances on the diagonal will be minimal when  $n_r = n_{max} = \max_r(n_1, n_2, n_3)$  for  $r \in 1, \dots, 3$ .

## 3 Secondary criteria

### 3.1 Precision

The previous section established that switching subclass changes  $V(\hat{\beta})$  so that one might reasonably write  $V_r(\hat{\beta})$  in recognition of this fact. The work above refers to the vector  $\text{diag}[V(\hat{\beta})]$  but we reduce this to a familiar and convenient scalar, namely the *total variance*, the trace of  $\text{tr}[V(\hat{\beta})]$  computed as

$$\text{tr}[V(\hat{\beta})] = \text{tr}[I^{-1}(\hat{\beta})] = \text{tr}[\sigma^2(X'X)^{-1}].$$

and use the total variance as our measure of precision.

### 3.2 Multi-collinearity

We note that in categorical data there will be some categories with low frequencies of occurrence and this may decrease the stability of the model by increasing *multi-collinearity*. We use  $K_r$  a generalisation of the condition number,  $K$ , of the design matrix  $X'X$  (Belsley, 2004) as a measure of multi-collinearity. Thus, we may investigate how the multi-collinearity of a model varies with  $r$  and the relationship between the pair  $(V_r, K_r)$  varies with  $r$ . The condition number of a square matrix,  $M$ , is defined as

$$K(M) = \sqrt{\lambda_{\max}/\lambda_{\min}} = \nu_{\max}/\nu_{\min},$$

where  $\lambda_{\max} = \text{maximum}(\lambda_j)$ ,  $\lambda_{\min} = \text{minimum}(\lambda_j)$ , and  $\lambda_j$ ,  $j = 1, 2, \dots, p$ , are the eigenvalues of  $M$  and the  $\nu$ s are the Singular Value Decomposition (SVD) numbers. Belsley (2004) derived the threshold values for  $\kappa(M = X'X)$  in the linear model and arrived at the threshold values of 10 and 30 indicating medium and serious degrees of multi-collinearity, respectively. Accordingly, we exploit Belsley's results viz:

$$K_r = K(M_r) = \sqrt{\lambda_{r,\max}/\lambda_{r,\min}} = \nu_{r,\max}/\nu_{r,\min},$$

for a matrix with categorical regressors and reference subclasses denoted by the subscript  $r$ .

In the linear model with two binary covariates we are able to find the explicit form for  $K_r$ .

$$K_r = K((X'X)_r) = \sqrt{\frac{R + 2\sqrt{Q} \cos\phi}{R - \sqrt{Q} (\cos\phi + \sqrt{3} \sin\phi)}} \quad (6)$$

where  $Q = \frac{2}{18}(n^2 + 4n_1^2 + 4n_2^2 + 3n_{12}^2 - nn_1 - nn_2 - n_1n_2)$  and  $0 < \phi < \pi/3$ .

## 4 Relationship between $V_r$ and $K_r$

As  $V_r$  increases  $K_r$  increases and the model becomes more unstable. We have examined a large number of cases in the linear model via a large simulation study (too detailed to reproduce here) covering different numbers of categories and sample sizes. In these studies the correlation between  $V_r$  and  $K_r$  is high almost always in the region of 0.95. It follows that in the linear model the most stable model is *always* found by minimizing the total variance, i.e., by choosing the reference subclass to be the subclass which contains the largest number of observations. This result is useful when dealing with sparse data.

## 5 Extension to GLMs

Here the quantities of interest generalize, but involve GLM specific weights and the findings are more complicated. In general we have a similar finding to that when the linear model obtains, i.e., a *good* strategy, one which reduces the multi-collinearity, is to select the most numerous subclass as the reference subclass. However, this it is not uniformly true and in the main paper we set forth details of the quantities involved and the conditions necessary to optimise the choice.

## 6 Summary

The stability of statistical models involving multiple categorical regressors can be improved by a judicious choice of the reference subclasses. This arises because  $V(\hat{\beta})$  varies with  $r$  and hence so does  $K_r$ . In general, in the linear model one should always choose the largest subclass as the reference subclass in order to minimise the multi-collinearity. In GLMs the position is similar, but only under certain conditions. It may be shown that optimal choice depends on a combination of the GLM weights and the disparity of the observed subclass distribution (Peng & MacKenze, 2014). Nevertheless, choosing the largest subclass is often a good strategy. The final paper gives more information on these matters. One consequence is that subject matter choices can be evaluated in terms of the pair  $(V_r, K_r)$ .

**Acknowledgments:** This work was supported by the SFI's ([www.sfi.ie](http://www.sfi.ie)) BIO-SI ([www3.ul.ie/bio-si](http://www3.ul.ie/bio-si)) research programme, grant number, **07MI012**.

## References

- Belsley, D. A., Kuh, E. and Welsch, R. E. (2004). Regression diagnostics: Identifying influential data and sources of collinearity, 1st edn. John Wiley and Sons.
- MacKenzie, G. and Peng, D.(2013). Interval-censored parametric regression survival models and the analysis of longitudinal trials. *Statistics in Medicine*: 32, 28042822.
- Peng, D. and MacKenzie, G. (2014). Discrepancy and choice of reference subclass in categorical regression models. In *Statistical Modelling in Biostatistics and Bioinformatics: selected papers*, Springer, Munich, ISBN 978-3-319-04578-8. (Forthcoming).