Bounds on the regression coefficients when a covariate is categorized

Kooreman, P.

Published in:
Communications in statistics: Part A: Theory and methods

Publication date:
1993

COMMUN. STATIST.—THEORY METH., 22(8); 2373-2380 (1993)

BOUNDS ON THE REGRESSION COEFFICIENTS WHEN A COVARIATE IS CATEGORIZED

Peter Kooreman

Wageningen University
P.O.B. 8060
6700 DA Wageningen
The Netherlands

Key Words and Phrases: Categorized Data, Linear Regression, Bounds, Matrix Differential Calculus.

ABSTRACT

This paper considers the problem of a linear regression model in which a continuous covariate is categorized. In such a case one can obtain bounds on the OLS-estimate which would be calculated if the complete information were available. These bounds give an indication of the information loss due to grouping and show what can be inferred from the data without additional assumptions.

1. INTRODUCTION

A problem that shows up frequently in empirical research is that a variable of interest is categorized (grouped), i.e. is only observed to fall in a certain

interval on a continuous scale, its actual value remaining unobserved. An example of such a variable is income. Consumers are often unable or unwilling to provide a precise income measure, and therefore income is usually recorded bracketwise.

Several "solutions" to the problem of a categorized covariate have been proposed in the literature, the simplest one being the use of interval midpoints as proxies for the true values. This approach, however, will generally yield inconsistent parameter estimates. Hsiao and Mountain (1985) have proposed an ingenious pseudo-instrumental variable method, but it requires a functional form assumption for the probability density of the continuous variable underlying the categorized covariate. One might also simply use dummy variables indicating in which interval the categorized covariate falls. However, this introduces artificial discontinuities in the effect of the categorized covariate and the estimation results are not easily translated into the parameter that would be estimated if the categorized covariate were completely observed.

This paper notes that the categorized data imply natural bounds on the OLS-estimate which would be calculated if the complete information were available. The bounds give an indication of the information loss due to grouping and show what can be inferred from the data without additional assumptions. They convey useful complementary information, especially if doubt exists about the validity of the additional assumptions that are necessary to obtain more precise estimates.

2. THE BOUNDS

Consider the model

$$y = X\beta+\epsilon \qquad (1)$$

where $X=(x_1,\ldots,x_K)$ is an (NxK)-matrix of deterministic covariates, β is a K-vector of unknown parameters, and ϵ and y are (Nx1)-vectors. It is assumed throughout that x_1, the first column of X, refers to the categorized covariate. So the elements of this column are unknown, but it is known that $L_{n1}\leq x_{n1}<U_{n1}$, $n=1,\ldots,N$. That is, for each x_{n1} the upper and lower bound of the interval in which it falls is known.

Obviously, one cannot calculate the OLS-estimate $b=(X'X)^{-1}X'y$ since the first column of X is unknown. However, for all $k=1,\ldots,K$, one can calculate the largest and smallest b_k that are still consistent with $L_{n1}\leq x_{n1}<U_{n1}$.

The upper bound on b_1 is the solution to the following problem

$$\max_{(x_{11},\ldots,x_{N1})} \; b_1=e_1'(X'X)^{-1}X'y \qquad (2)$$
$$\text{s.t.} \qquad L_1\leq x_1\leq U_1$$

where $e_1=(1,0,\ldots 0)'$, $L_1=(L_{11},\ldots,L_{N1})'$ and $U_1=(U_{11},\ldots,U_{N1})'$.

The problem entails maximization of a continuously differentiable function on a compact set, so that a solution exists. Define $\lambda=(\lambda_1,\ldots,\lambda_N)'$ and $\mu=(\mu_1,\ldots,\mu_N)'$. Then the Lagrangean function associated with problem (2) can be written as

$$\mathcal{L} = b_1(x_{11},\ldots,x_{N1}) + \lambda'(x_1-L_1) + \mu'(U_1-x_1) \qquad (3)$$

From Magnus and Neudecker (1988, p. 307)

$$db=[(X'X)^{-1}\otimes u'-b'\otimes(X'X)^{-1}X']dvecX \qquad (4)$$

where $u=y-Xb$. [In the notation $db=A.dc$ the element a_{ij}

of the matrix A contains the partial derivative of the i-th element of b with respect to the j-th element of c.] Since $b_i=e_i'b$, we obtain

$$\begin{aligned} db_i &= e_i'db \\ &= [(e_i'\otimes 1)((X'X)^{-1}\otimes u') \\ &\quad -(1\otimes e_i')(b'\otimes(X'X)^{-1}X')]dvecX \\ &= [(e_i'(X'X)^{-1}\otimes u')-b'\otimes e_i'(X'X)^{-1}X']dvecX. \end{aligned} \tag{5}$$

Let $f=(X'X)^{-1}e_i$. Then

$$db_i=(f\otimes u-b\otimes Xf)'dvecX \tag{6}$$

However, we require the differential of b_i when only the elements of the first column of X are perturbed. Therefore

$$dX=(dx_i)\iota' \tag{7}$$

(ι is Kx1 unit vector). Hence

$$db_i=(f_iu-bXf)'dx_i, \tag{8}$$

where f_i is the first element of f. The Kuhn-Tucker conditions that characterize the solution can now be written as

$$\begin{aligned} &(f_iu-bXf)+\lambda-\mu=0 \\ &\Lambda(x_i-L_i)=0 \\ &M(U_i-x_i)=0 \\ &L_i\leq x_i\leq U_i,\ \lambda\geq 0,\ \mu\geq 0 \end{aligned} \tag{9}$$

where Λ and M are diagonal matrices with λ and μ on the diagonal, respectively. The bounds on the other coefficients are characterized in the same manner.

The generalization to the case of two or more categorized variables is conceptually straightforward. For example, if the first two columns of X are categorized variables, then b_1 in equation (2) should be maximized with respect to $(x_{11},\ldots,x_{N1},x_{12}\ldots,x_{N2})$ subject to $L_{n1}\leq x_{n1}<U_{n1}$ and $L_{n2}\leq x_{n2}<U_{n2}$, $n=1,\ldots,N$.

3. CALCULATING THE BOUNDS

To illustrate the bounds two samples were drawn for the categorized covariate x_1 from a lognormal distribution [log $x_{n1}\sim N(10;1)$ with right truncation at 120000] with sizes 100 and 1000. Next, x_{n1} was categorized into the following intervals: (0,15000], (15000,30000], (30000,45000], (45000,60000] and (60000,120000]. The response variable is defined as

$$y_n = 20000+\beta_1 x_{n1}+\epsilon_n, \qquad n=1,\ldots,N$$

where ϵ_n are drawings from $N(0;40000^2)$.

Let b^{PROX} be the estimator obtained by using some number M_j as a proxy for the true value of x_{n1} if x_{n1} falls in interval $(A_{j-1},A_j]$, $j=1,\ldots,J$. It is straightforward to verify that

$$plim\ b_1^{PROX} = \beta_1 . \frac{\sum_{j=1}^{J} p_j(\mu_j-\bar{\mu})(M_j-\bar{M})}{\sum_{j=1}^{J} p_j(M_j-\bar{M})^2}, \qquad (10)$$

where $p_j=Pr(A_{j-1}<x_{n1}\leq A_j)$, $\mu_j=E(x_{n1}\,|\,A_{j-1}<x_{n1}\leq A_j)$, $\bar{\mu}=\Sigma_j p_j\mu_j=Ex_{n1}$ and $\bar{M}=\Sigma_j p_j M_j$. Hence, in general b^{PROX} is consistent only

TABLE I.

		N=100	N=1000
$\beta_1=1$	b_1	0.979	0.997
	bounds	[0.502;1.946]	[0.482;2.089]
$\beta_1=0.1$	b_1	0.079	0.094
	bounds	[-0.377;0.444]	[-0.389;0.846]
$\beta_1=10$	b_1	9.979	9.997
	bounds	[6.766;18.96]	[6.806;18.67]
$\beta_1=-1$	b_1	-1.021	-1.003
	bounds	[-2.177;-0.587]	[-2.067;-0.481]

if $M_j=E(x_{n1}|A_{j-1}<x_{n1}\leq A_j)$, $j=1,\ldots,J$. If the M_j's are taken to be the midpoints of the intervals, plim $b_1^{PROX}=0.862\beta_1$ in the present example. The Hsiao and Mountain approach estimates β consistently if the distributional assumption with respect to x_1 is correct.

The bounds were calculated on a mainframe VAX using the NAG-Library routine E04KBF, which maximizes an arbitrary differentiable function subject to upper and/or lower bounds on the variables, using a Newton type method with analytical first derivatives. As a check, the bounds were also calculated in GAUSS386 (PC) using a steepest-ascent type algorithm. Identical results were obtained in all cases. The results appeared to be insensitive to various choices of starting values. For four different values of β_1 table 1 reports the OLS-estimate b_1 (using the complete information) and the bounds (using the categorized information).

4. DISCUSSION

First note that there is no systematic relationship between the bound width and the sample size.

Tighter bounds can only be obtained if the number of intervals in which x_1 is categorized gets larger and their widths smaller. Only for $\beta_1=0.1$ the bounds include zero. In the other cases the bounds indicate that one can be confident about the sign of b_1.

Nevertheless, the bounds may seem wide and, in fact, they represent a 'worst case'. Consider the first row of table 1, N=1000, for example. The bounds state that if the complete information would have been available, the OLS-estimate of β_1 based on the complete information could neither have been smaller than 0.482 nor larger than 2.089. These values are achieved when for 1000 and 737 observations, respectively, the categorized variable is at one of the two boundaries of its interval. It will often be unlikely that the true distribution of x_1 has such a shape. More precise estimates are obtained if one of the approaches described in the introduction is employed. Of course, the additional precision then stems from the additional assumptions that are invoked. The bounds provide useful complementary information and allow readers to make a tradeoff between the benefits of additional precision and the price that has to be paid for it in terms of additional assumptions.

One of the questions for additional research using the approach presented here is drawing statistical inference with respect to β. In principle, inference may be approached in a similar way as estimation. For example, considering the (estimated) covariance matrix

of the OLS estimator as a function of x_1, each diagonal element of this matrix can be maximized and minimized subject to $L_1 \leq x_1 \leq U_1$ to find the smallest and largest variances consistent with the bounds.

ACKNOWLEDGEMENTS

I thank J.S. Cramer, Denzil Fiebig and two anonymous referees for helpful comments.

BIBLIOGRAPHY

Hsiao, C. and D. Mountain (1985), "Estimating the Short-Run Income Elasticity of Demand for Electricity by Using Cross-Sectional Categorized Data", *J. Amer. Stat. Assoc.*, 80 (No. 390, Applications), pp. 259-265.

Magnus, J. and H. Neudecker (1988), *Matrix Differential Calculus; with Applications to Statistics and Econometrics*, John Wiley and Sons.

Received August 1992; Revised February 1993