Journal of Economic and Social Measurement 30 (2005) 279–293
IOS Press

The decomposition of inter-group differences in a logit model: Extending the Oaxaca-Blinder approach with an application to school enrolment in India

Vani K. Borooah[a] and Sriya Iyer[b,*]
[a]*School of Economics and Politics, University of Ulster, Northern Ireland*
[b]*Faculty of Economics and St. Catharine's College, University of Cambridge, Cambridge, UK*

This paper suggests a method of decomposing differences in inter-group probabilities from a logit model and shows how it can be related to similar decompositions derived from a Oaxaca-Blinder framework. In so doing, it offers a solution to a problem, embedded within the Oaxaca-Blinder decomposition, relating to the appropriate choice of common coefficient vectors with which to evaluate the different attribute vectors. The decomposition method also shows how pair-wise comparisons of groups might be conducted in the presence of more than two groups, without discarding the information on groups excluded from the comparison. The proposed method is applied to inter-group differences in schooling participation in India and the results are compared with the Oaxaca-Blinder method. The decomposition is applied specifically to inter-group differences in the enrolment of boys at school in India.

Keywords: Logit regression, school enrolment, India, decomposition methods

1. Introduction

The Oaxaca [9] and Blinder [4] method of decomposing group differences in means as the sum of a "coefficients" contribution and an "attributes" contribution is, arguably, the most widely used decomposition technique in economics. This method has been extended from its original setting within regression analysis, to explaining group differences in probabilities derived from models of discrete choice with a binary dependent variable and estimated using logit and probit methods [1–3, 7,8]. However, there are two constricting aspects of this decomposition that are often overlooked.

First, the Oaxaca-Blinder decomposition (and its extension) are formulated for situations in which the sample is subdivided into two mutually exclusive and (collectively exhaustive) groups, such as, for example, men and women. Then, one may decompose the difference in, for example, average wages between men and women – or the difference between men and women in their average probabilities of being employed in a particular occupational class – into two parts: the first due to gender

*Corresponding author: Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, UK. Tel.: +44 1223 335257; Fax: +44 1223 335475; E-mail: Sriya.Iyer@econ.cam.ac.uk.

0747-9662/05/$17.00 © 2005 – IOS Press and the authors. All rights reserved

differences in the coefficient vectors and the other owing to gender differences in attributes.

The attribute contribution is computed by asking what the average male-female difference in wages would have been if the difference in attributes between men and women had been evaluated using a common coefficient vector. The critical question though is what should be this common coefficient vector? Typically, two separate computations of the attribute contribution are provided using, respectively, the male and the female coefficient vectors as the common vector. But there is a problem here: the estimate of the degree of "gender discrimination" – defined as the total difference less the attribute contribution – may be different (perhaps, considerably) between the two computations. The decomposition as it is formulated, offers no solution to this conundrum.

The second difficulty is that in many situations one may wish to subdivide the population into more than two groups (for example, Hispanic, Black, White). The Oaxaca-Blinder decomposition may be applied to such situations through the pair-wise comparison of groups, ignoring groups excluded from a particular comparison. So, for example, one may apply the Oaxaca-Blinder decomposition to the difference in mean wages between Whites and Blacks, ignoring the presence of Hispanics; or to the difference in mean wages between Blacks and Hispanics, ignoring the presence of Whites. The problem with this procedure is that by discarding data on the third group, it, in effect, reduces the tripartite division of the sample into a binary one. And the problem is aggravated if the population is to be subdivided into more groups.

The decomposition proposed in this paper shows how pair-wise comparisons may be conducted without discarding data on groups not involved in the comparisons. The essential idea is to ask what the mean outcome would be if *everyone's* attributes were, successively, evaluated according to the coefficient vector of a particular group (everyone was evaluated using White, Black, and Hispanic coefficients). Since the *only* factor that would alter between these evaluations is the group coefficient vector according to which the evaluations were conducted, one may identify the difference in outcomes between these evaluations as being generated *entirely* by inter-group coefficient differences. The difference between the observed mean outcome for a group and the mean outcome for the entire sample, if everyone was evaluated using that group's coefficients, may be regarded as being due to attribute differences between that group and the other groups.

The remainder of this paper formalises these ideas by showing how the decomposition method proposed can be related to the familiar Oaxaca-Blinder method. In so doing, it offers a solution to a problem, embedded within the Oaxaca-Blinder decomposition, relating to the appropriate choice of a common coefficient vector with which to evaluate the different attribute vectors. The decomposition method proposed suggests how pair-wise comparisons of groups might be conducted in the presence of more than two groups, without discarding information on groups excluded from the comparison. The proposed method is compared with the Oaxaca-Blinder method when both are applied to inter-group differences in schooling participation in India.

2. The econometric framework

There are N children (indexed, $i = 1 \ldots N$) who can be placed in K mutually exclusive and collectively exhaustive groups, $k = 1..K$, each group containing N_k children. Define the variable *ENR*$_i$ such that *ENR*$_i = 1$, if the child is enrolled at school, *ENR*$_i = 0$, if the child is not enrolled. Then, under a logit model, the likelihood of a child, from group k, being enrolled in school is:

$$\Pr(ENR_i = 1) = \frac{\exp(\mathbf{X}_i^k \beta^k)}{1 + \exp(\mathbf{X}_i^k \beta^k)} = F(\mathbf{X}_i^k \hat{\beta}^k) \tag{1}$$

where: $\mathbf{X}_i^k = \{X_{ij},\, j = 1...J\}$ represents the vector of observations, for child i of group k, on J variables which determine the likelihood of the child being enrolled at school, and $\hat{\beta}^k = \left\{\beta_j^k,\, j = 1...J\right\}$ is the associated vector of coefficient estimates for children belonging to group k.

The average probability of a child from group k being enrolled at school – which is also the mean enrolment rate for the group – is:

$$\bar{ENR}^k = \bar{P}(\mathbf{X}_i^k, \hat{\beta}^k) = N_k^{-1} \sum_{i=1}^{N_k} F(\mathbf{X}_i^k \hat{\beta}^k) \tag{2}$$

Now for any two groups, say Hindu ($k = H$) and Muslim ($k = M$):

$$\begin{aligned} \bar{ENR}^H - \bar{ENR}^M = [\bar{P}(\mathbf{X}_i^M, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^M)] + [\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) \\ -\bar{P}(\mathbf{X}_i^M, \hat{\beta}^H)] \end{aligned} \tag{3}$$

Alternatively:

$$\begin{aligned} \bar{ENR}^H - \bar{ENR}^M = [\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^H, \hat{\beta}^M)] + [\bar{P}(\mathbf{X}_i^H, \hat{\beta}^M) \\ -\bar{P}(\mathbf{X}_i^M, \hat{\beta}^M)] \end{aligned} \tag{4}$$

The first term in square brackets, in Eqs (3) and (4), represents the "coefficients effect": it is the difference in average enrolment rates between Hindu and Muslim children resulting from applying different coefficients vector to a given vector of attribute values. The second term in square brackets in Eqs (3) and (4) represents the "attributes effect": it is the difference in average enrolment rates between Hindu and Muslim children resulting from inter-group differences in attributes, when these attributes are evaluated using a common coefficient vector.

So for example, in Eq. (3), the difference in sample means is decomposed by asking what the average school enrolment rates for Muslim children would have been, *had they been treated as Hindus*; in Eq. (4), it is decomposed by asking what the average school enrolment rates for Hindu children would have been, *had they been treated*

as Muslim. In other words, the common coefficient vector used in computing the attribute effect is, for Eq. (3), the Hindu vector and, for Eq. (4), the Muslim vector.

The problem with this method of decomposition – call it the "Oaxaca-Blinder" logistic decomposition – is that Eqs (3) and (4) are separate equations: the decomposition is anchored either by treating Muslims as Hindus (as in Eq. (3)) or Hindus as Muslims (as in Eq. (4)). We now propose a decomposition method which combines the elements of Eqs (3) and (4) into a single decomposition formula.

3. An extension of the Oaxaca-Blinder decomposition method

For the purposes of exposition, suppose there are three groups: Hindus ($k = H$); Muslims ($k = M$); and *Dalits*[1] ($k = D$) whose population shares are, respectively, θ^H, θ^M and θ^D. Define the quantities $\bar{P}^r$ (for $r = H, M, D$) as:

$$\bar{P}^r = N^{-1} \sum_k \sum_{i=1}^{N_k} \left[\frac{\exp(\mathbf{X}_i^k \beta^r)}{1 + \exp(\mathbf{X}_i^k \beta^r)} \right] = N^{-1} \sum_k \sum_{i=1}^{N_k} F[(\mathbf{X}_i^k \beta^r)] \tag{5}$$

Then $\bar{P}^r$ is the average probability of enrolment computed over *all* the children in the sample when their individual attribute vectors (the $\mathbf{X}_i^k$) are *all* evaluated using the coefficient vector of group r (β^r); equivalently, $\bar{P}^r$ is the average probability of enrolment, computed over the entire sample, *when all the children are treated as belonging to group r*. Hereafter, $\bar{P}^r$ is referred to as the group r *synthetic* probability of school enrolment. Suppose for two groups, r and s, $\bar{P}^r > \bar{P}^s$. Then the difference in synthetic probabilities, $\bar{P}^r - \bar{P}^s$, represents the greater advantage to children from belonging to group r compared to belonging to group s. This difference is identified as the "coefficients effect" because it is entirely the consequence of a given set of attributes (that of all the N children in the sample) evaluated using different coefficient vectors.

The difference between the observed, average enrolment rate of Hindu children ($E\bar{N}R^H$) and the Hindu 'synthetic probability' of school enrolment ($\bar{P}^H$), may be regarded as being due to attribute differences between Hindu children and children from the other two groups, Muslim and Dalit. More formally:

$$E\bar{N}R^H - \bar{P}^H = \left\{ \bar{P}(\mathbf{X}_i^H, \hat{\beta}^H - N^{-1} \left(\sum_{i=1}^{N_H} F[\mathbf{X}_i^H \hat{\beta}^H] + \sum_{i=1}^{N_D} F[\mathbf{X}_i^D \hat{\beta}^H] \right. \right.$$
$$\left. \left. + \sum_{i=1}^{N_D} F[\mathbf{X}_i^D \hat{\beta}^H] \right) \right\}$$

[1] Those castes and tribes – also known as Scheduled Castes/Tribes – recognised by the Indian Constitution in 1947 as deserving special recognition in respect of education, employment and political representation.

$$
\begin{aligned}
&= \left\{\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \theta^H \bar{P}(\mathbf{X}_{fi}^H, \hat{\beta}^H) - \theta^M \bar{P}(\mathbf{X}_i^M, \hat{\beta}^H) - \theta^D \bar{P}(\mathbf{X}_i^D, \hat{\beta}^H)\right\} \\
&= \left\{\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \theta^H \bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \theta^M \bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \theta^D \bar{P}(\mathbf{X}_i^H, \hat{\beta}^H)\right\} \\
&\quad +\theta^M[\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^H)] + \theta^D[\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^D, \hat{\beta}^H)] \\
&= \theta^M[\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^H)] + \theta^D[\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^D, \hat{\beta}^H)] \qquad (6)
\end{aligned}
$$

Equation (6) says that the difference between the observed enrolment rate of Hindu children and the Hindu synthetic probability of enrolment ($E\bar{N}R^H - \bar{P}^H$) is the weighted sum of the difference in probabilities arising from Hindu and Muslim attributes, and of Hindu and Dalit attributes, being evaluated using the Hindu coefficient vector estimates (respectively, $\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^H)$ and $\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^D, \hat{\beta}^H)$), the weights ($\theta^M$ and θ^D) being, respectively, the proportion of Muslims and Dalits in the sample. Similarly:

$$
\begin{aligned}
&E\bar{N}R^M - \bar{P}^M = \theta^H[\bar{P}(\mathbf{X}_{\mathbf{i}}^M, \hat{\beta}^M) - \bar{P}(\mathbf{X}_i^H, \hat{\beta}^M)] + \theta^D[\bar{P}(\mathbf{X}_i^M, \hat{\beta}^M) \\
&-\bar{P}(\mathbf{X}_{\mathbf{i}}^D, \hat{\beta}^M)] = -\theta^H[\bar{P}(\mathbf{X}_i^H, \hat{\beta}^M) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^M)] - \theta^D[\bar{P}(\mathbf{X}_i^D, \hat{\beta}^M) \\
&-\bar{P}(\mathbf{X}_i^M, \hat{\beta}^M)] \qquad (7)
\end{aligned}
$$

Then, using Eqs (6) and (7), the difference in mean enrolment rates between Hindus and Muslims may be written as:

$$
\begin{aligned}
&E\bar{N}R^H - E\bar{N}R^M = E\bar{N}R^H - \bar{P}^H + \bar{P}^H - E\bar{N}R^M + \bar{P}^M - \bar{P}^M \\
&= (\bar{P}^H - \bar{P}^M) + [(E\bar{N}R^H - \bar{P}^H) - (E\bar{N}R^M - \bar{P}^M)] \\
&= (\bar{P}^H - \bar{P}^M) + \theta^M\left\{\bar{P}(\mathbf{X}_{\mathbf{i}}^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^H)\right\} \\
&\quad +\theta^H\left\{\bar{P}(\mathbf{X}_{\mathbf{i}}^H, \hat{\beta}^M) - \bar{P}(\mathbf{X}_i^M, \beta^M)\right\} \\
&\quad +\theta^D\left\{\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^D, \hat{\beta}^H)\right\} \\
&\quad +\theta^D\left\{\bar{P}(\mathbf{X}_i^D, \hat{\beta}^M) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^M)\right\} \\
&= \Omega + \Lambda
\end{aligned} \qquad (8)
$$

As the decomposition formula in Eq. (8) shows, the difference between Hindu and Muslim children in their mean enrolment rates can be written as the sum of a coefficients effect (Ω) and an attributes effect (Λ). The coefficients effect is the difference between the Hindu and Muslim synthetic probabilities ($\Omega = \bar{P}^H - \bar{P}^M$) and the attributes effect is:

$$
\begin{aligned}
\Lambda &= \theta^M\left\{\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_{\mathbf{i}}^M, \hat{\beta}^H)\right\} \\
&\quad +\theta^H\left\{\bar{P}(\mathbf{X}_i^H, \hat{\beta}^M) - \bar{P}(\mathbf{X}_i^M, \beta^M)\right\} \\
&\quad +\theta^D\left\{\bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^D, \hat{\beta}^H)\right\} \\
&\quad +\theta^D\left\{\bar{P}(\mathbf{X}_i^D, \hat{\beta}^M) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^M)\right\}
\end{aligned} \qquad (9)
$$

The expression for Λ, in Eq. (9), shows that the components of the *overall* attribute effect are:

(i) Differences in attributes between Muslims and Hindus, *evaluated at Hindu coefficients* (weight: proportion of Muslims in the sample, θ^M)
(ii) Differences in attributes between Muslims and Hindus, *evaluated at Muslim coefficients* (weight: proportion of Hindus in the sample, θ^H)
(iii) Differences in attributes between Hindus and Dalits, *evaluated at Hindu coefficients* (weight: proportion of Dalits in the sample, θ^D)
(iv) Differences in attributes between Muslims and Dalits, *evaluated at Muslim coefficients* (weight: proportion of Dalits in the sample, θ^D)

When there are only two groups, $\theta^D = 0$, $\theta^M + \theta^H = 1$, Eq. (8) becomes:

$$\begin{aligned} E\bar{N}R^H - E\bar{N}R^M = (\bar{P}^H - \bar{P}^M) + \theta^M \left\{ \bar{P}(\mathbf{X}_i^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^H) \right\} \\ + \theta^H \left\{ \bar{P}(\mathbf{X}_i^H, \hat{\beta}^M) - \bar{P}(\mathbf{X}_i^M, \beta^M) \right\} \end{aligned} \tag{10}$$

Comparing the decomposition formula of Eq. (10) to the Oaxaca-Blinder decomposition of Eqs (3) and (4) shows that the "attribute effect" terms of Eqs (3) and (4) – respectively, $\bar{P}(\mathbf{X}_\mathbf{i}^H, \hat{\beta}^H) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}^H)$ and $\bar{P}(\mathbf{X}_i^H, \hat{\beta}^M) - \bar{P}(\mathbf{X}_i^M, \beta^M)$ - *both* enter decomposition formula of Eq. (10), appropriately weighted by the population shares of the two groups. Conversely, if θ^M and θ^H are simply regarded as weights then Eqs (3) and (4) can be obtained from Eq. (10) by setting θ^M or θ^H to zero.

With three groups, there are, as Eq. (8) shows, two further "attribute effect" terms to be considered. The first of these involves Dalits and Hindus and it is reflected in the change in the average probability of enrolment when the Hindu and Dalit sub-samples are evaluated using Hindu coefficients; the second term involves Dalits and Muslims and it is reflected in the change in the average probability of enrolment when the Muslim and Dalit sub-samples are evaluated using Muslim coefficients. Each of these terms is weighted by the population share of Dalits. Since the calculation of $\bar{P}^H$ and $\bar{P}^M$ involved *all* the children in the sample, these additional residual terms adjust for the fact that this included Dalit children.

If Hindus and Muslims had the same vector of coefficient estimates, so that $\hat{\beta}^H = \hat{\beta}^M$, then $\bar{P}^H = \bar{P}^M$, Eq. (10) becomes:

$$E\bar{N}R^H - E\bar{N}R^M = \bar{P}(\mathbf{X}_i^H, \hat{\beta}) - \bar{P}(\mathbf{X}_i^M, \hat{\beta}) \tag{11}$$

implying that the difference between Hindus and Muslims in the proportions of children enrolled at school would be entirely due to differences between them in attributes.

It is possible to further decompose the "coefficients effect", using an indicator variable which serves as one of the explanatory variables in the logit Eq. (8). Suppose that the region in which the children live is one such variable; if there are M regions,

Table 1
Selected Data for School Enrolments by Group: Boys Aged 6–14

	Hindus (*10, 178 boys*)	*Muslims* (*2,300 boys*)	*Dalits* (*7,367 boys*)
% boys enrolled	84	68	70
% boys enrolled: Central	79	59	61
% boys enrolled: South	86	91	80
% boys enrolled: West	91	83	81
% boys enrolled: East	86	62	73
% boys enrolled: North	93	68	81
% boys enrolled: both parents literate	96	93	92
% boys enrolled: both parents illiterate	70	50	58
% boys enrolled: cultivator father	85	67	69
% boys enrolled: labourer father	74	57	64
% boys enrolled: non-manual father	89	74	80

Children whose both parents were present in the household.
Source: NCAER Survey.

indexed, *m=1...M*, such that N_m children live in region m, of whom N_m^k are from group k, then $\bar{P}^r$ (of Eq. (5)) can be rewritten as:

$$\bar{P}^r = \sum_{m=1}^{M} \mu_m N_m^{-1} \sum_{k=1}^{K} \sum_{i=1}^{N_m^k} \bar{P}(\mathbf{X}_i^k, \beta_m^r) = \sum_{m=1}^{M} \mu_m \bar{P}_m^r \tag{12}$$

where: $\mu_m = (N_m/N)$ is the proportion of children in the sample who live in region m; β_m^r is the coefficient vector of group r in region m; and $\bar{P}_m^r$ is the average probability of enrolment in region m (*m=1...M*), if *all* the children in region m were treated as belonging to group r.

Then, from Eq. (12), for any two groups r and s:

$$\bar{P}^r - \bar{P}^s = \sum_{m=1}^{M} \mu_m (\bar{P}_m^r - \bar{P}_m^s) \tag{13}$$

and $\mu_m(\bar{P}_m^r - \bar{P}_m^s)/(\bar{P}^r - \bar{P}^s)$ is the proportionate contribution that region m makes to the overall coefficients effect. Note that $\bar{P}_m^r = \bar{P}_m^s$ if $\beta_m^r = \beta_m^s$ and that $\bar{P}^r = \bar{P}^s$ if $\beta_m^r = \beta_m^s$ for all *m=1...M*.

4. An application

Consider first the logit equation for school enrolment specified as:

$$\log\left(\frac{\Pr(ENR_i = 1)}{1 - \Pr(ENR_i = 1)}\right) = \sum_{j=1}^{J} \beta_j X_{ij} + \sum_{j=1}^{J} \beta_j^M (M_i \times X_{ij}) \tag{14}$$

$$+\sum_{j=1}^{J}\beta_j^D(D_i \times X_{ij})$$

in which: X_{ij} is the value of j^{th} (*j=1...J*) determining variable for child i (*i=1...N*); β_j is the 'Hindu coefficient' associated with the j^{th} (*j=1...J*) determining variable; and β_j^M and β_j^D are the *changes* to this coefficient from being, respectively, Muslim and Dalit.

The econometric estimates are based on unit record data from the 1993–94 Human Development Survey of India [10]. This survey encompasses 33,000 *rural* households – 195,000 individuals – which were spread over 1,765 villages, in 195 districts, drawn from 16 states of India[2]. Equation (14) was estimated on data for 19,845 boys aged 6–14. Table 1 shows the salient features of the relevant data and the estimation results are shown in Table 2.

There were some variables for which the coefficients were significantly different between the groups: the β_j^M and/or the β_j^D were significantly different from zero implying that, associated with these variables, there were additional effects from being Muslim or Dalit. Such variables are clearly identified in Table 2. Some of these effects were regional: Muslim and Dalit boys living in the Central region had *ceteris paribus* a lower likelihood of being enrolled at school than their Hindu counterparts. Some of these effects related to parental occupation: in particular, *ceteris paribus* Dalit boys with fathers who were cultivators had a lower likelihood of being enrolled at school than their Hindu and Muslim counterparts. Some of these effects related to institutional infrastructure: the presence of *anganwadis* (or informal 'courtyard classrooms') in villages did more to boost the school enrolment rates of Muslim, relative to Hindu, boys.

Table 3 shows the results from the 'Oaxaca-Blinder' logistic decompositions. These show that, of the Hindu-Muslim difference in the mean enrolment rate of boys, 64% – when Muslims were treated as Hindus (Eq. (3)) – and 48% – when Hindus were treated as Muslims (Eq. (4)) – could be attributed to coefficient differences: these percentages reflected the contribution of the 'coefficents effect' towards explaining inter-group differences in mean enrolment rates.

The coefficients effect played a much smaller role in explaining differences in mean enrolment rates between Hindus and Dalits: respectively, 43% of the difference in the Hindu-Dalit enrolment rate for boys could be explained by inter-group coefficient differences, when Dalits were treated as Hindus (Eq. (3)); when Hindus were treated as Dalits (Eq. (4)), the corresponding figure was 36%.

[2]This survey – commissioned by the Indian Planning Commission and funded by a consortium of United Nations agencies – was carried out by the National Council of Applied Economic Research (NCAER) over January-June 1994 and most of the data from the survey pertains to the year prior to the survey, that is to 1993–94. Details of the survey – hereafter referred to as the NCAER Survey – are to be found in Shariff (1999), though some of the salient features of data from the NCAER Survey, insofar as they are relevant to this study, are described in the Data Appendix to this paper.

Table 2
Logit Estimates of the School Enrolment Equation: 19,845 Boys, 6–14 years

Determining Variables	*Coefficient Estimate* (*z value*)	*Marginal Probabilities*
Muslim	−0.4075898	−0.160
	(5.16)	
Dalit	−0.7991797	−0.033
	(2.49)	
Central	−0.5079733	−0.100
	(9.91)	
South	–	–
West	–	–
East	−0.6417705	−0.072
	(4.08)	
Household Income	1.002299	0.0003
	(3.01)	
Father educated: low	2.792598	0.128
	(20.84)	
Mother educated: low*	2.634748	0.113
	(11.44)	
Father educated: medium**	2.921865	0.121
	(14.48)	
Mother educated: medium**	2.114656	0.087
	(5.14)	
Father educated: high**	3.890858	0.148
	(16.71)	
Mother educated: high***	2.1909003	0.089
	(4.01)	
Father cultivator	1.474474	0.056
	(6.37)	
Father labourer	–	
Father non-manual	1.550021	0.060
	(7.45)	
Mother Cultivator	–	
Mother labourer	−0.7691638	−0.041
	(3.06)	
Mother non-manual	−0.5848008	−0.092
	(3.22)	
No anganwadi in village	−0.8018316	−0.032
	(5.07)	
No primary school in village	–	
No middle school within 2 km	−0.8358139	−0.027
	(4.21)	
Number of Siblings	−0.8985882	−0.016
	(7.20)	
Additional Effects of Muslims		
Central	−0.4962503	
	(4.10)	
East	−0.3896603	
	(4.80)	
Father educated: medium	1.734144	
	(2.70)	

Table 2, continued

Determining Variables	*Coefficient Estimate* (*z value*)	*Marginal Probabilities*
Mother labourer	1.795181	
	(2.62)	
Mother non-manual	6.466559	
	(2.41)	
Anganwadi	1.739127	
	(4.40)	
Middle School	1.508577	
	(3.55)	
Number of Siblings	1.091813	
	(2.56)	
Additional Effects of Dalits		
Central	−0.8562861	
	(1.71)	
East	−0.7160941	
	(2.38)	
Father cultivator	−0.8704603	
	(1.77)	
Mother labourer	1.221465	
	(1.88)	

Figures in parentheses are z-values and coefficients are shown in terms of 'odds-ratios'.

Table 3
The Decomposition of Inter-Group Differences in the Proportion of Boys Enrolled at School: "Oaxaca-Blinder type" Logistic Decomposition

	Sample Average	*Group s treated as group r*		*Group r treated as group s*	
	$E\bar{N}R^r - E\bar{N}R^s$	$\bar{P}(\mathbf{X}_i^s, \hat{\beta}^r)$ $-\bar{P}(\mathbf{X}_i^s, \hat{\beta}^s)$	$\bar{P}(\mathbf{X}_i^r, \hat{\beta}^r)$ $-\bar{P}(\mathbf{X}_i^s, \hat{\beta}^r)$	$\bar{P}(\mathbf{X}_i^r, \hat{\beta}^r)$ $-\bar{P}(\mathbf{X}_i^r, \hat{\beta}^s)$	$\bar{P}(\mathbf{X}_i^r, \hat{\beta}^s)$ $-\bar{P}(\mathbf{X}_i^s, \hat{\beta}^)$
$r =$ **Hindu**	$0.843 - 0.675$	$0.782 - 0.675$	$0.843 - 0.782$	$0.843 - 0.763$	$0.763 - 0.675$
$s =$ **Muslim**	$=$ **0.168**	$=$ **0.107**	$=$ **0.061**	$=$ **0.080**	$=$ **0.088**
$r =$ **Hindu**	$0.843 - 0.698$	$0.760 - 0.698$	$0.843 - 0.760$	$0.843 - 0.791$	$0.791 - 0.698$
$s =$ **Dalit**	$=$ **0.145**	$=$ **0.062**	$=$ **0.083**	$=$ **0.052**	$=$ **0.093**
$r =$ **Dalit**	$0.698 - 0.675$	$0.713 - 0.675$	$0.698 - 0.713$	$0.698 - 0.660$	$0.660 - 0.675$
$s =$ **Muslim**	$=$ **0.023**	$=$ **0.038**	$=$ **−0.015**	$=$ **0.038**	$=$ **−0.015**

Although differences between Dalits and Muslims, in the mean enrolment rates, were not as marked as between each of these groups and the Hindus, this lack of difference concealed considerable differences between Dalits and Muslims in terms of enrolment-enhancing attributes and attitudes. Broadly speaking, Muslims were better endowed with enrolment-enhancing attributes and qualitative evidence from the survey showed that Dalits had a more positive attitude towards school participation. And this is seen clearly when Muslim attributes were evaluated using Dalit coefficients: the mean enrolment of Muslim boys *rose* from 68% to 71% (Table 3, right panel); on the other hand, when Dalit attributes were evaluated using Muslim coefficients, the mean enrolment of Dalit boys *fell* from 70% to 66% (Table 3, left panel).

Table 4
The Decomposition of Inter-Group Differences in the Proportion of Boys Enrolled at School: "Recycled Proportions" Logistic Decomposition

	Difference in Average Enrolment Rates	*The Coefficents* Effect*	*The Attribute** Effect*
	$E\bar{N}R^r - E\bar{N}R^s$	$\bar{P}^r - \bar{P}^s$	$(E\bar{N}R^r - E\bar{N}R^s) - (\bar{P}^r - \bar{P}^s)$
$r =$**Hindu** $s =$ **Muslim**	$0.843 - 0.675 =$ **0.168**	$0.805 - 0.714 =$ **0.091**	$0.168 - 0.091 =$ **0.077**
$r =$ **Hindu** $s =$ **Dalit**	$0.843 - 0.698 =$ **0.145**	$0.805 - 0.748 =$ **0.057**	$0.145 - 0.057 =$ **0.088**
$r =$ **Dalit** $s =$ **Muslim**	$0.698 - 0.675 =$ **0.023**	$0.748 - 0.714 =$ **0.034**	$0.023 - 0.034 =$ **−0.011**

* Difference in the average probabilities of school enrolment when *all* persons were assumed to belong to group r against *all* persons belonging to group s.
** Calculated as the weighted sum of the individual Blinder-Oaxaca attribute effects (Eq. (8)).

Table 3 also makes clear that the proportion of the difference in mean enrolment rates of boys, between Hindus and Muslims that could be ascribed to inter-group coefficient differences, varied markedly (64%–48%) depending upon whether Muslims were treated as Hindus (equation (3)) or Hindus were treated as Muslims (Eq. (3)). A comparison of Hindu and Dalit enrolment rates showed a similar variation (43%–36%).

The decomposition method suggested in this paper, as discussed earlier, overcomes this difficulty. Table 4 shows that 54% of the difference between the Hindu and Muslim average enrolment rates, and 39% of the difference between Hindu and Dalit enrolment rates, for boys could be ascribed to the "coefficients effect".

To what extent does the coefficient associated with a particular attribute contribute to the overall "coefficients effect"? Table 5 (using Eq. (13)) shows that, between Hindus and Muslims, 65% of the overall coefficients effect in the enrolment rate of boys was contributed by the Central region and 27% was contributed by the Eastern region with the percentage contributions of the 'high enrolment rate regions' of the South, the West and the North being negligible. A similar story could be told with respect to Dalits. This suggests that inter-group 'attitudinal' differences towards the education of boys were, by and large, associated with the poorer regions of India where the school enrolment rate was low.

5. The statistical significance of coefficient and attribute effects

There is an important issue which, while we do not address directly in this paper, is nevertheless, an important one to flag. This has to do with the statistical significance of the coefficients associated with the individual variables in the econometric equations and (the consequent) statistical significance of the individual variable contributions to the difference in means.

Table 5
The Regional Contributions to the all-India "Coefficents Effect": Boys

	Central	*South*	*West*	*East*	*North*	*All-India*
Hindus v Muslims:						
$\mu_m(\bar{P}_m^H - \bar{P}_m^M)$	0.059	0.003	0.002	0.024	0.003	0.091
Percentage contribution	65	3	2	27	3	100
Hindus v Dalits						
$\mu_m(\bar{P}_m^H - \bar{P}_m^M)$	0.036	0.004	0.002	0.012	0.003	0.057
Percentage contribution	63	7	4	21	5	100

The percentage distribution of the 19,845 boys in the sample between the regions were: Central (46.8), South (17.3), West (11.5), East (13.9); North (10.6).

Yun [12] proposed a method to evaluate the contributions of the individual variables to the overall difference in group means: i.e. to coefficients *and* the attributes effect. By contrast, Nielsen [8] and this paper analysed the individual contributions to only the coefficients effect while Even and MacPherson [5,6] carried out a similar analysis for only the attributes effect. Yun [13] then used asymptotic theory to propose significance tests for the aggregate coefficients and attributes effects and also for the individual contributions to the coefficients and attributes effects.

Although we do not address this question in this paper – regarding it, in the context of the methodology proposed, as an area for future development – Yun's [12, 13] work has important implications for studies (such as our own) which involve the decomposition of differences in the first moment. This is because some of the coefficients associated with particular explanatory variables may not be – indeed, inevitably, are not – statistically significant. Consequently, this raises the parallel question of whether they make a significant contribution to the attributes effect and, at an extreme, whether the attributes effect, considered in its entirety, is itself significantly different from zero.

6. Conclusion

This paper suggested a method of decomposing differences in inter-group probabilities from a logit model and has shown how it might be viewed as an extension of decompositions derived from the Oaxaca-Blinder framework. In so doing, it has offered a solution to a problem, embedded within the Oaxaca-Blinder decomposition, relating to the appropriate choice of a common coefficient vector with which to evaluate the different attribute vectors. This decomposition method also shows how pair-wise comparisons of groups might be conducted in the presence of more than two groups, without discarding information on groups excluded from the comparison. This is a particularly important consideration when applying decomposition methods to investigating inter-group differences in economic circumstances in pluralistic societies.

The decomposition technique was applied to examine inter-group differences in India in the enrolment of boys at school. This gave rise to two broad conclusions:

first, that Muslims in India were better endowed with enrolment-enhancing attributes but that Dalits had a more positive attitude towards school enrolment. Second, that inter-group 'attitudinal' differences towards the education of boys were predominantly associated with the poorer regions of India where the overall rate of school enrolment is very low. These decomposition methods, therefore, also have important implications for the causes of difference among ethnically diverse populations in poor countries.

Acknowledgements

We are grateful to the National Council of Applied Economic Research (NCAER), New Delhi for providing us with the unit record data from its 1993–94 Human Development Survey on which this study is based. Thanks are due especially to the Editor and two anonymous referees for several useful comments and suggestions. Iyer also acknowledges support and funding from The British Academy.

Data Appendix

The data used for estimating Eq. (13) were obtained from the NCAER survey, referred to earlier. All regression models were estimated using the software package STATA Release 5. The salient features of this data are set out in this section. The data from the NCAER survey are organised as a number of 'reference' files, with each file focusing on specific subgroups of individuals. However, the fact that in every file an individual was identified by a household number and, then, by an identity number within the household, meant that the 'reference' files could be joined – as described below – to form larger files.

So, for example, the schooling equations were estimated on data from the 'individual' file. This file, as the name suggests, gave information on the 194,473 individuals in the sample with particular reference to their educational attainments, amongst other information about them. From this file, data on the school enrolment of each male child aged 6–14 were extracted (the variable ENR) and associated with this information was data on: the educational attainments and occupation of the boy's father and/or mother; the income and size of the household to which the boy belonged; the state, district and village in which he lived; his caste/tribe (Dalit, non-Dalit); his religion; the number of his siblings etc. The equation relating to school enrolment was estimated on data from the NCAER Survey's 'Individual' file', described above, for boys between the ages of 6–14 (inclusive) who had both parents living in the household: this yielded a total of 19,845 observations.

Another file – the 'village file' – contained data relating to the existence of infrastructure in, and around, each of the 1,765 villages over which the survey was con-

ducted. This file gave information as to whether *inter alia* a village: had *anganwadis* [3], primary schools, middle schools and high schools and, if it did not, what was the nature of access to such institutions. The village file could be joined to the individual file so that for each individual (say, boy between 6–14) there was information not just on the his schooling outcome and on his family and household circumstances but also on the quality of the educational facilities – and general infrastructure – in the village in which he lived.

The sample of children was distinguished by three *mutually exclusive* subgroups: Dalits; Muslims; and Hindus. In effect, the Hindu/Muslim/Dalit distinction made in this paper is a distinction between: non-Dalit Hindus; Muslims; and Dalit Hindus. These subgroups are, hereafter, referred to as 'groups'. Because of the small number of Christians and persons of 'other' religions in the Survey, the analysis reported in this paper was confined to Hindus, Muslims and Dalits.

The Survey contained information for each of sixteen states. In this study, the states were aggregated to form five regions: the *Central* region consisting of Bihar, Madhya Pradesh, Rajasthan and Uttar Pradesh; the *South* consisting of Andhra Pradesh, Karnataka, Kerala and Tamilnadu; the *West* consisting of Maharashtra and Gujarat; the *East* consisting of Assam, Bengal and Orissa; and the *North* consisting of Haryana, Himachal Pradesh and Punjab.

References

[1] D.H. Blackaby, S. Drinkwater, D.G. Leslie and P.D. Murphy, A Picture of Male Unemployment Among Britain's Ethnic Minorities, *Scottish Journal of Political Economy* **44** (1997), 182–197.
[2] D.H. Blackaby, D.G. Leslie, P.D. Murphy and N.C. O'Leary, The Ethnic wage Gap and Employment Differentials in the 1990s: Evidence for Britain, *Economics Letters* **58** (1998), 97–103.
[3] D.H. Blackaby, D.G. Leslie, P.D. Murphy and N.C. O'Leary, Unemployment Among Britain's Ethnic Minorities, *The Manchester School* **67** (1999), 1–20.
[4] A.S. Blinder, Wage Discrimination: Reduced Form and Structural Estimates, *Journal of Human Resources* **8** (1973), 436–455.
[5] W.E. Even and D.A. MacPherson, Plant Size and the Decline of Unionism, *Economics Letters* **32** (1990), 393–398.
[6] W.E. Even and D.A. MacPherson, The Decline of Private Sector Unionism and the Gender Wage Gap, *Journal of Human Resources* **28** (1993), 279–296.
[7] J. Gomulka and N. Stern, The Employment of Married Women in the United Kingdom, 1970–83, *Economica* **57** (1990), 171–199.
[8] H.S. Nielsen, Discrimination and Detailed Decomposition in a Logit Model, *Economics Letters* **61** (1998), 115–120.

[3] *Anganwadis* are village-based early childhood development centres. They were devised in the early 1970s as a baseline village health centre, their role being to: provide state government-funded food supplements to pregnant women and children under five; to work as an immunization outreach agent; to provide information about nutrition and balanced feeding, and to provide vitamin supplements; to run adolescents girls' and women's groups; and to monitor the growth, and promote the educational development of, children in a village.

[9] R. Oaxaca, Male-Female Wage Differentials in Urban Labor Markets, *International Economic Review* **14** (1973), 693–709.
[10] A. Shariff, *India Human Development Report* (1999), New Delhi: Oxford University Press.
[11] StataCorp, Stata Statistical Software: Release 5.0, 1997, College Station, TX: Stata Corporation.
[12] M.-S. Yun, Decomposing Differences in the First Moment, *Economics Letters* **82** (2004), 275–280.
[13] M.-S. Yun, *Hypothesis Tests When Decomposing Differences in the First Moment*, 2004, Department of Economics, Tulane University (mimeo).