The Decomposition of Inter-Group Differences in a Logit Model: Extending the Oaxaca-Blinder Approach with an Application to School Enrolment in India

Borooah, Vani and Iyer, Sriya

University of Ulster, University of Cambridge

2005

The Decomposition of Inter-Group Differences in a Logit Model: Extending the Oaxaca-Blinder Approach with an Application to School Enrolment in India[Ψ]

Vani K. Borooah*
University of Ulster, Northern Ireland

Sriya Iyer#
University of Cambridge, England

June 2003

Abstract

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 pairwise 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-community differences in the enrolment of boys at school in India.

JEL Classification: J7 Keywords: Logit; School Enrolment; India

---

[Ψ] 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. The second author acknowledges support and funding from the British Academy. Needless to say, the usual disclaimer applies.

*Corresponding author: School of Economics and Politics, University of Ulster, Newtownabbey BT37 0QB, Northern Ireland UK; Tel: +44-28-9096-1357; Fax: +44-28-9096-1356; E-mail: VK.Borooah@ulst.ac.uk

[#] Faculty of Economics and Politics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DD; E-mail: Sriya.Iyer@econ.cam.ac.uk

1. Introduction

The Oaxaca (1973) and Blinder (1973) method of decomposing group differences in means into a "discrimination" and a "characteristics" component 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/probit methods (Gomulka and Stern, 1990; Blackaby *et. al*., 1997, 1998,1999; Nielsen, 1998). However, there are two constricting aspects of this decomposition and of its extension to logit/probit models, 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 "managerial" position – into two parts, one due to gender differences in the coefficient vectors and one due to gender differences in the attribute (or variable) vectors.

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 vary (perhaps, greatly) between the two computations. The decomposition as it stands, 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/probabilities between Whites and Blacks, ignoring the presence of Hispanics; or to the difference in mean wages/probabilities between Blacks and Hispanics, ignoring the presence of Whites. The problem with this procedure is that by discarding data on the third group, in effect it reduces the tripartite division of the sample into a binary one. And the problem is intensified if the population may be subdivided into many more groups.

The decomposition proposed here 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 (wages; probability of an event) would be if *everyone* (White, Black, Hispanic) was, successively, treated as belonging *exclusively* to a particular group (all-White; all-Black; all-Hispanic). Since the *only* factor that is altered between these experiments is the group to which the individuals are assigned, one may identify the difference in outcomes between these experiments as being generated *entirely* by group membership. The difference between the observed outcome for a group (mean wage/probability for Whites) and its “experimental outcome” (mean wage/probability, computed over the entire sample, if everyone was treated as being White) may then, intuitively, be assigned to attribute differences between the particular group and the other groups.

The following pages formalise these ideas by showing how the decomposition method proposed relates 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 pairwise 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…N*) who can be placed in *K* mutually exclusive and collectively exhaustive groups (hereafter referred to as 'communities'), *k=1..K,* each community 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 community *k*, being enrolled in school is:

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

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

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

$$E\bar{N}R^k = \bar{P}(\mathbf{X_i^k}, \hat{\boldsymbol{\beta}}^\mathbf{k}) = {N_k}^{-1}\sum_{i=1}^{N_k} F(\mathbf{X_i^k}\hat{\boldsymbol{\beta}}^\mathbf{k}) \tag{2}$$

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

$$E\bar{N}R^H - E\bar{N}R^M = [\bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}^\mathbf{H}) - \bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}^\mathbf{M})] + [\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^\mathbf{H}) - \bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}^\mathbf{H})] \tag{3}$$

Alternatively:

$$E\bar{N}R^H - E\bar{N}R^M = [\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^\mathbf{H}) - \bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^\mathbf{M})] + [\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^\mathbf{M}) - \bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}^\mathbf{M})] \tag{4}$$

The first term in square brackets, in equations (3) and (4), represents the "response effect": it is the difference in average enrolment rates between Hindu and Muslim children resulting from inter-community differences in responses (as exemplified by differences in the coefficient vectors) to a given vector of attribute values. The second term in square brackets in equations

(3) and (4) represents the "attributes effect": it is the difference in average enrolment rates between Hindu and Muslim children resulting from inter-community differences in attributes, when these attributes are evaluated using a common coefficient vector.

So for example, in equation (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 equation (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 equation (3), the Hindu vector and, for equation (4), the Muslim vector.

The problem with this method of decomposition – call it the "Oaxaca-Blinder" logistic decomposition – is that equations (3) and (4) are separate equations: the decomposition is anchored either by treating Muslims as Hindus (as in equation (3)) or Hindus as Muslims (as in equation (4)). In the section 2.1, a method of decomposition is proposed which combines the elements of equations (3) and (4) into a single decomposition formula.

2.1 *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,k=H,M,D*) as:

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

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* ($\mathbf{\beta^r}$); equivalently, $\bar{P}^r$ is the average probability of enrolment, computed over the entire sample, *when all*

[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.

the children are treated as belonging to community r. Hereafter, $\overline{P}^r$ is referred to as the community *r synthetic* probability of school enrolment. For any two communities, the difference between them in their synthetic probabilities, $\overline{P}^r - \overline{P}^s$, represents the difference in the advantage to children, as measured by the average probability of being enrolled at school, between belonging to community *r* and to community *s.* This difference is identified as the "response effect" because it is entirely the consequence of differences between communities *r* and *s* in their responses to a given vector of attributes.

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

$$\begin{aligned}
&E\overline{N}R^H - \overline{P}^H = \left\{ \overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - N^{-1}\left(\sum_{i=1}^{N_H} F[\mathbf{X_i^H}\hat{\boldsymbol\beta}^\mathbf{H}] + \sum_{i=1}^{N_M} F[\mathbf{X_i^M}\hat{\boldsymbol\beta}^\mathbf{H}] + \sum_{i=1}^{N_D} F[\mathbf{X_i^D}\hat{\boldsymbol\beta}^\mathbf{H}] \right) \right\} \\
&= \left\{ \overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \theta^H \overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \theta^M \overline{P}(\mathbf{X_i^M}, \hat{\boldsymbol\beta}^\mathbf{H}) - \theta^D \overline{P}(\mathbf{X_i^D}, \hat{\boldsymbol\beta}^\mathbf{H}) \right\} \\
&= \left\{ \overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \theta^H \overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \theta^M \overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \theta^D \overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) \right\} \\
&+ \theta^M [\overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \overline{P}(\mathbf{X_i^M}, \hat{\boldsymbol\beta}^\mathbf{H})] + \theta^D [\overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \overline{P}(\mathbf{X_i^D}, \hat{\boldsymbol\beta}^\mathbf{H})] \\
&= \theta^M [\overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \overline{P}(\mathbf{X_i^M}, \hat{\boldsymbol\beta}^\mathbf{H})] + \theta^D [\overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{H}) - \overline{P}(\mathbf{X_i^D}, \hat{\boldsymbol\beta}^\mathbf{H})]
\end{aligned} \tag{6}$$

Equation (6) says that the difference between the observed enrolment rate of Hindu children and the Hindu synthetic probability of enrolment 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, the weights being, respectively, the proportion of Muslims and Dalits in the sample. Similarly:

$$\begin{aligned}
&E\overline{N}R^M - \overline{P}^M = \theta^H [\overline{P}(\mathbf{X_i^M}, \hat{\boldsymbol\beta}^\mathbf{M}) - \overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{M})] + \theta^D [\overline{P}(\mathbf{X_i^M}, \hat{\boldsymbol\beta}^\mathbf{M}) - \overline{P}(\mathbf{X_i^D}, \hat{\boldsymbol\beta}^\mathbf{M})] \\
&= -\theta^H [\overline{P}(\mathbf{X_i^H}, \hat{\boldsymbol\beta}^\mathbf{M}) - \overline{P}(\mathbf{X_i^M}, \hat{\boldsymbol\beta}^\mathbf{M})] - \theta^D [\overline{P}(\mathbf{X_i^D}, \hat{\boldsymbol\beta}^\mathbf{M}) - \overline{P}(\mathbf{X_i^M}, \hat{\boldsymbol\beta}^\mathbf{M})]
\end{aligned} \tag{7}$$

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

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

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

$$
\begin{aligned}
\Lambda = & \theta^{M}\left\{\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^{\mathbf{H}}) - \bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}^{\mathbf{H}})\right\} + \theta^{H}\left\{\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^{\mathbf{M}}) - \bar{P}(\mathbf{X_i^M}, \boldsymbol{\beta}^{\mathbf{M}})\right\} \\
& +\theta^{D}\left\{\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^{\mathbf{H}}) - \bar{P}(\mathbf{X_i^D}, \hat{\boldsymbol{\beta}}^{\mathbf{H}})\right\} + \theta^{D}\left\{\bar{P}(\mathbf{X_i^D}, \hat{\boldsymbol{\beta}}^{\mathbf{M}}) - \bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}^{\mathbf{M}})\right\}
\end{aligned}
$$

The expression for Λ, above, shows that the components of the *aggregate* 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, $\theta^{D} = 0$, $\theta^{M} + \theta^{H} = 1$ and equation (8) becomes:

$$
\begin{aligned}
& E\bar{N}R^{H} - E\bar{N}R^{M} = (\bar{P}^{H} - \bar{P}^{M}) \\
& +\theta^{M}\left\{\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^{\mathbf{H}}) - \bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}^{\mathbf{H}})\right\} + \theta^{H}\left\{\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^{\mathbf{M}}) - \bar{P}(\mathbf{X_i^M}, \boldsymbol{\beta}^{\mathbf{M}})\right\}
\end{aligned}
\qquad (9)
$$

Comparing the decomposition formula of equation (9) – call it the "recycled proportions" logistic decomposition - to that in equations (3) and (4) shows that the "attribute effect" terms of equations (3) and (4) - respectively,

$\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^{\mathbf{H}}) - \bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}^{\mathbf{H}})$ and $\bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}^{\mathbf{M}}) - \bar{P}(\mathbf{X_i^M}, \boldsymbol{\beta}^{\mathbf{M}})$ - *both* enter the decomposition formula of equation (8), appropriately weighted by the population shares of the two groups. Conversely, if θ^M and θ^H are simply regarded as weights then equations (3) and (4) can be obtained from equation (9) by setting θ^M or θ^H to zero.

With three groups, there are, as equation (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 subsamples 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 subsamples 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 $\hat{\boldsymbol{\beta}}^{\mathbf{H}} = \hat{\boldsymbol{\beta}}^{\mathbf{M}}$, then $\bar{P}^H = \bar{P}^M$ and equation (8) becomes:

$$E\bar{N}R^H - E\bar{N}R^M = \bar{P}(\mathbf{X_i^H}, \hat{\boldsymbol{\beta}}) - \bar{P}(\mathbf{X_i^M}, \hat{\boldsymbol{\beta}}) \qquad (10)$$

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 "response effect", using an indicator variable which serves as one of the explanatory variables in the logit equation (Nielsen, 1998). Suppose that the region in which the children live is one such variable; if there are *M* regions, indexed, *m=1…M*, such that N_m children live in region *m*, of whom N_m^k are from community *k,* then $\bar{P}^r$ (of equation (5)) can be rewritten as:

$$\overline{P}^r = \sum_{m=1}^{M} \mu_m N_m^{-1} \sum_{k=1}^{K} \sum_{i=1}^{N_m^k} \overline{P}(\mathbf{X_i^k}, \boldsymbol{\beta}_\mathbf{m}^\mathbf{r}) = \sum_{m=1}^{M} \mu_m \overline{P}_m^r \qquad (11)$$

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

Then, from equation (11), for any two communities *r* and *s*:

$$\overline{P}^r - \overline{P}^s = \sum_{m=1}^{M} \mu_m (\overline{P}_m^r - \overline{P}_m^s) \qquad (12)$$

and $\mu_m (\overline{P}_m^r - \overline{P}_m^s)/(\overline{P}^r - \overline{P}^s)$ is the proportionate contribution that region *m* makes to the overall response effect. Note that $\overline{P}_m^r = \overline{P}_m^s$ if $\boldsymbol{\beta}_\mathbf{m}^\mathbf{r} = \boldsymbol{\beta}_\mathbf{m}^\mathbf{s}$ and that $\overline{P}^r = \overline{P}^s$ if $\boldsymbol{\beta}_\mathbf{m}^\mathbf{r} = \boldsymbol{\beta}_\mathbf{m}^\mathbf{s}$ for all *m=1…M*.

3. An Application

Consider first the logit equation for school enrolment specified as:

$$\log\left(\frac{\Pr(ENR_i = 1)}{1 - \Pr(ENR_i = 1)}\right) = \sum_{j=1}^{J} \beta_j X_{ij} + \sum_{j=1}^{J} \beta_j^M (M_i \times X_{ij}) + \sum_{j=1}^{J} \beta_j^D (D_i \times X_{ij}) \qquad (13)$$

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 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 (Sharif 1999). 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 (13) was

[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.

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 communities: 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 (equation (3)) - and 48% - when Hindus were treated as Muslims (equation (4)) - could be attributed to coefficient differences: these percentages reflected the contribution of the 'response effect' towards explaining inter-community differences in mean enrolment rates.

The response 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-community coefficient differences, when Dalits were treated as Hindus (equation (3)); when Hindus were treated as Dalits (equation (4)), the corresponding figure was 36%.

Although differences between Dalits and Muslims, in the mean enrolment rates, were not as marked as between each of these communities 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 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-community coefficient differences, varied markedly (64%-48%) depending upon whether Muslims were treated as Hindus (equation (3)) or Hindus were treated as Muslims (equation (4)). 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 "response effect".

To what extent does the "attribute effect" contribute to the "response effect"? Table 5 (using equation (12)) shows that 65% of the overall response effect, between Hindus and Muslims, 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-community 'attitudinal' differences towards the education of boys were, by and large, associated with the poorer regions of India where the overall rates of school enrolment was low.

4. Conclusion

This paper has 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 pairwise 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-community 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-community '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.

References

Blackaby, D.H., Drinkwater, S., Leslie, D.G., Murphy, P.D. (1997), "A Picture of Male Unemployment Among Britain's Ethnic Minorities", Scottish Journal of Political Economy, vol. 44, pp. 182-197.

Blackaby, D.H., Leslie, D.G., Murphy, P.D. and O'Leary, N.C. (1998), "The Ethnic wage Gap and Employment Differentials in the 1990s: Evidence for Britain", Economics Letters, vol. 58, pp. 97-103.

Blackaby, D.H., Leslie, D.G., Murphy, P.D. and O'Leary, N.C. (1999), "Unemployment Among Britain's Ethnic Minorities", The Manchester School, vol. 67, pp. 1-20.

Blinder, A.S. (1973), "Wage Discrimination: Reduced Form and Structural Estimates", Journal of Human Resources, vol. 8, pp. 436-455.

Gomulka, J. and Stern, N. (1990), "The Employment of Married Women in the United Kingdom, 1970-83", Economica, vol. 57, pp. 171-199.

Nielsen, H.S. (1998), "Discrimination and Detailed Decomposition in a Logit Model", Economics Letters, vol. 61, pp. 115-20.

Oaxaca, R. (1973), "Male-Female Wage Differentials in Urban Labor Markets", International Economic Review, vol. 14, pp. 693-709.

Shariff, A. (1999), India Human Development Report, New Delhi: Oxford University Press.

Data Appendix

The data used for estimating equation (13) were obtained from the NCAER survey, referred to earlier. 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[3]. 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 conducted. This file gave information as to whether *inter alia* a village: had *anganwadis*[4], 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

---

[3] Needless to say, the file also contained other information on the individuals.

[4] *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.

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 'communities'. 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.

Table 1

Selected Data for School Enrolments by Community: 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

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)	
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-Community Differences in the Proportion of Boys Enrolled at School: "Oaxaca-Blinder type" Logistic Decomposition

	Sample Average	*Community s treated as community r*		*Community r treated as community s*	
	$E\bar{N}R^r - E\bar{N}R^s$	$\bar{P}(\mathbf{X}_i^s, \hat{\boldsymbol{\beta}}^r) - \bar{P}(\mathbf{X}_i^s, \hat{\boldsymbol{\beta}}^s)$	$\bar{P}(\mathbf{X}_i^r, \hat{\boldsymbol{\beta}}^r) - \bar{P}(\mathbf{X}_i^s, \hat{\boldsymbol{\beta}}^r)$	$\bar{P}(\mathbf{X}_i^r, \hat{\boldsymbol{\beta}}^r) - \bar{P}(\mathbf{X}_i^r, \hat{\boldsymbol{\beta}}^s)$	$\bar{P}(\mathbf{X}_i^r, \hat{\boldsymbol{\beta}}^s) - \bar{P}(\mathbf{X}_i^s, \hat{\boldsymbol{\beta}}^s)$
***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**

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

	Difference in Average Enrolment Rates	*The Response 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**	0.843-0.675=	0.805-0.714=	0.168-0.091=
***s*=Muslim**	**0.168**	**0.091**	**0.077**
***r*=Hindu**	0.843-0.698=	0.805-0.748=	0.145-0.057=
***s*=Dalit**	**0.145**	**0.057**	**0.088**
***r*=Dalit**	0.698-0.675=	0.748-0.714=	0.023-0.034=
***s*=Muslim**	**0.023**	**0.034**	**-0.011**

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

Table 5
The Regional Contributions to the all-India "Response 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).