Extending the Oaxaca-Blinder Decomposition to the Independent Double Hurdle Model: With Application to Parental Spending on Education in Malawi

Mussa, Richard

Economics Department, Chancellor College, University of Malawi

2014

Extending the Oaxaca-Blinder Decomposition to the Independent Double Hurdle Model: With Application to Parental Spending on Education in Malawi

Richard Mussa*

December 18, 2014

Abstract

The study develops the Blinder-Oaxaca decomposition technique for the independent double hurdle model. The proposed decomposition is done at the aggregate level. Using the Second Malawi Integrated Household Survey (IHS2), the paper applies the proposed decomposition to explain the rural-urban difference in parental spending on own primary school children. The results show that at least 66% of the expenditure differential is explained by differences in characteristics between rural and urban households.

Keywords: Double Hurdle; Decomposition; Blinder-Oaxaca; Malawi

1 Introduction

The Blinder-Oaxaca decomposition, independently proposed by Oaxaca (1973) and Blinder (1973), is a popular tool for identifying and quantifying differences in economic outcomes such as earnings, income, and schooling between two groups or periods. The economic outcomes for two groups for example male-female, rural-urban or two time periods are decomposed at their mean. This intergroup or interperiod mean level decomposition is used for linear models. However, the standard Blinder-Oaxaca decomposition method cannot be used to decompose nonlinear models such as the probit, logit, tobit, and the independent double hurdle. Estimating a linear model when in fact the correct model is a nonlinear one can lead to biased and inconsistent coefficients, and this may in turn lead to an inaccurate decomposition.

For nonlinear models; Fairlie (1999, 2005) has proposed the Blinder-Oaxaca decomposition for logit and probit models, Bauer and Sinning (2008, 2010) have proposed an extension of the same for tobit models. The aim of this paper is to propose an extension of the Blinder-Oaxaca decomposition to the independent double hurdle model. The

*Department of Economics, Chancellor College, University of Malawi, Box 280, Zomba, Malawi, rimussa@yahoo.co.uk

proposed decomposition is done at the aggregate level. The aggregated decomposition isolates the gap in an economic outcome between two groups into a characteristic effect, which is a part of the gap explained by differences in social-economic characteristics, and a coefficient effect which is the part of the gap which is due to differences in coefficients. We illustrate the proposed decomposition by using Malawian data to analyze the rural-urban gap in parental spending on own primary school children. To the best of our knowledge this paper is the first to apply the Blinder-Oaxaca decomposition to examine differences in household expenditure.

2 Extending the Oaxaca-Blinder Decomposition to the Double Hurdle Model

One underlying feature of expenditure data is that it contains excess zeros, and the choice of a statistical technique used to deal with the zeros is important, as an inappropriate treatment of zeros can lead to biased and inconsistent estimates (Greene, 1981)[1]. The tobit model (Tobin, 1958) has been widely used to model outcomes which have excess zeros. The tobit model is derived from an individual optimization problem and views zeros as corner solution outcomes. The major drawback of the tobit model is that it assumes that the same stochastic process determines both the extensive and intensive margins, that is, the decision whether or not to spend (participation decision) and how much (expenditure decision), are treated as the same. This assumption is restrictive. A model which corrects this limitation of the tobit model is the double hurdle model (DH hereafter).

The DH model, originally formulated by Cragg (1971), assumes that households make two decisions with regard to spending, each of which is determined by a different underlying stochastic process (Blundell and Meghir, 1987). The double hurdle model estimated separately for two groups of households, $m = (U, R)$, with U and R denoting urban and rural households respectively, is formally specified as follows (Jones, 1989);

The participation equation (the first hurdle) is given as;

$$\begin{aligned} D^*_{im} &= Z'_{im}\alpha + \varepsilon_{im} \\ D_{im} &= \begin{cases} 1 \; if \; D^*_{im} > 0 \\ 0 \; otherwise \end{cases} \end{aligned} \tag{1}$$

[1] Although the paper presents the decomposition in terms of the rural-urban expenditure difference; the proposed decomposition can also be used for identifying the causes of regional, cross-country, time period, or other differences in an outcome variable. Besides, the outcome variable need not necessarily be expenditure; the proposed decomposition applies to all cases which require the estimation of the independent double hurdle.

The expenditure equation (the second hurdle) is given as follows;

$$\begin{aligned} Y_{im}^{*} &= X_{im}'\beta + \nu_{im} \\ Y_{im}^{**} &= \max(0, Y_{im}^{*}) \end{aligned} \tag{2}$$

Observed expenditure(Y_{im});

$$Y_{im} = D_{im} Y_{im}^{**} \tag{3}$$

where; D_{im}^{*} is a latent variable describing the household's decision to participate (spend or not) on children's education, Y_{im}^{*} is a latent variable describing household expenditure on children's education, Z_{im}' is a vector of variables explaining the participation decision, X_{im}' is a vector of variables explaining the expenditure decision. The parameter vectors are α_m, β_m assumed to be linear. $\varepsilon_{im}, \nu_{im}$ are independent random errors with the following properties; $\varepsilon_{im} \sim N(0,1)$ and $\nu_{im} \sim N(0, \sigma^2)$, and i denotes household. The assumption of independence between ε_{im} and ν_{im} is quite common when using the DH (see for example Mauldin et al. (2001), Jensen and Yen (1996), Su and Yen (1996)). The alternative would be to assume that the errors are dependent. However, Smith (2003) shows that there is little statistical information to support the estimation of a DH with dependent errors even when dependence exists.

To derive the Blinder-Oaxaca decomposition for the independent DH; consider the DH as expressed in equation 3. We want to decompose the gap in the average value of the dependent variable for rural and urban households, $\triangle^{DH} = E(Y_U) - E(Y_R)$, by using the following sample counterpart $\hat{\triangle}^{DH} = \bar{Y}_U - \bar{Y}_R$. The sample average expenditure share for group m is given as $\bar{Y}_m = \frac{\sum_{i=1}^{N_m} \hat{Y}_{im}}{N_m}$; where N_m is the sample size for group m. The "hat" denotes sample estimates. The Blinder-Oaxaca decomposition of the independent DH similar to that for the Tobit by Bauer and Sinning (2008, 2010) is expressed in terms of unconditional expectations of the dependent variable (Y_{im}). The unconditional expectation for the two groups is expressed as follows;

$$E(Y_{im}) = \Pr(Y_{im} > 0) E(Y_{im} | Y_{im} > 0) \tag{4}$$

Where the probability of expenditure is given by;

$$\Pr(Y_{im} > 0) = \Phi\left(Z_{im}'\alpha_m\right) \Phi\left(\frac{X_{im}'\beta_m}{\sigma_m}\right) \tag{5}$$

and the conditional expectation of Y_{im} is expressed as;

$$E(Y_{im}|Y_{im}>0) = X'_{im}\beta_m + \frac{\sigma_m \phi\left(\frac{X'_{im}\beta_m}{\sigma_m}\right)}{\Phi\left(\frac{X'_{im}\beta_m}{\sigma_m}\right)} \tag{6}$$

Three things need to be noted about equation 4. Firstly, the unconditional expectation $E(Y_{im})$ is not equal to $E(X_{im})'\beta_m$ as is the case in linear models on which the standard Blinder-Oaxaca decomposition is based[2]. As discussed earlier, imposing a linear model on a dependent variable with excess zeros leads to biased and inconsistent coefficients, and therefore using coefficients from the linear model would give a misleading decomposition as well. Secondly, the unconditional expectation is not equal to that of Tobit as it has another censoring mechanism, $\Phi\left(Z'_{im}\alpha_m\right)$ which represents participation; this means that we cannot use the Blinder-Oaxaca decomposition for Tobit models as developed by Bauer and Sinning (2008, 2010). Finally, equation 4 shows that the unconditional expectation has the standard error of the error term of the expenditure equation, σ_m. This may affect the magnitude of the decomposition, and therefore has to be included in the decomposition. As a result, there are several possible decompositions of the mean difference $\triangle^{DH}$, depending on which σ_m is used in the counterfactual part of the decomposition.

We therefore derive two possible decompositions for the independent DH[3]:

$$\begin{aligned}\triangle \underset{R_1}{^{DH}} = &\left[E_{\alpha_U,\beta_U,\sigma_U}\left(Y_{iU}\right) - E_{\alpha_U,\beta_U,\sigma_R}\left(Y_{iR}\right)\right] \\ &+ \left[E_{\alpha_U,\beta_U,\sigma_R}\left(Y_{iR}\right) - E_{\alpha_R,\beta_R,\sigma_R}\left(Y_{iR}\right)\right]\end{aligned} \tag{7}$$

and

$$\begin{aligned}\triangle \underset{U_1}{^{DH}} = &\left[E_{\alpha_U,\beta_U,\sigma_U}\left(Y_{iU}\right) - E_{\alpha_U,\beta_U,\sigma_U}\left(Y_{iR}\right)\right] \\ &+ \left[E_{\alpha_U,\beta_U,\sigma_U}\left(Y_{iR}\right) - E_{\alpha_R,\beta_R,\sigma_R}\left(Y_{iR}\right)\right]\end{aligned} \tag{8}$$

Where $E_{\alpha_m,\beta_m,\sigma_m}\left(Y_{im}\right)$ denotes the unconditional expectation of Y_{im} evaluated at the parameter vectors α_m, β_m and the error standard error σ_m. The difference between the

[2] Assuming a linear model $Y_{im} = X'_i\beta + \nu_i$ for illustration; the standard Blinder-Oaxaca decomposition is based on the property of linear models with an intercept that the mean of a dependent variable is equal to the mean of the regressors evaluated at their respective estimated coefficients i.e. $\bar{Y}_{im} = \bar{X}_{im}\hat{\beta}_m$. Hence, the standard Blinder-Oaxaca decomposition is given as; $\bar{Y}_U - \bar{Y}_R = \left(\bar{X}_U\hat{\beta}_U - \bar{X}_R\hat{\beta}_R\right) = \left(\bar{X}_U - \bar{X}_R\right)\hat{\beta}_U + \left(\hat{\beta}_U - \hat{\beta}_R\right)\bar{X}_R$.
Where the "overbars" denote sample means and the "hats" denote sample estimates.

[3] These two possibilities are similar to that of Bauer and Sinning (2005) for the Tobit.

two decompositions is that equation 7 treats the standard error as part of the variables while equation 8 treats it as part of the coefficients.

The above decompositions use the urban coefficients in the counterfactual; this implies that if there was no gap in average expenditure share, the expenditure profile of the urban would prevail. We can alternatively use the rural coefficients; this implies that if there was no gap in average expenditure, the expenditure structure of the rural areas would exist. When the rural coefficients are used the two possibilities are written as[4]:

$$\begin{aligned} \Delta \ {}^{DH}_{U_2} = & \left[E_{\alpha_R, \beta_R, \sigma_U} (Y_{iU}) - E_{\alpha_R, \beta_R, \sigma_R} (Y_{iR}) \right] \\ & + \left[E_{\alpha_U, \beta_U, \sigma_U} (Y_{iR}) - E_{\alpha_R, \beta_R, \sigma_U} (Y_{iR}) \right] \end{aligned} \tag{9}$$

and

$$\begin{aligned} \Delta \ {}^{DH}_{R_2} = & \left[E_{\alpha_R, \beta_R, \sigma_R} (Y_{iU}) - E_{\alpha_R, \beta_R, \sigma_R} (Y_{iR}) \right] \\ & + \left[E_{\alpha_U, \beta_U, \sigma_U} (Y_{iR}) - E_{\alpha_R, \beta_R, \sigma_R} (Y_{iR}) \right] \end{aligned} \tag{10}$$

The first term in the decompositions (equations 7-10) captures part of the average expenditure share gap between the urban and rural households attributable to differences in covariates. This is the *characteristic effect*. This basically is the part of the gap in average expenditure share between the two groups of households assuming that both types had the same coefficients (behavior) but different endowments. Thus, this is a part of the gap explained by differences in characteristics. The last term in equations 7-10, measures the difference in average expenditure between the two groups which is due to differences in coefficients. This is the *coefficient effect*. It is part of the gap which is unexplained by the differences in characteristics. Essentially, it is part of the gap assuming that urban and rural households had the same characteristics but different coefficients. We interpret the coefficient effect as part of the gap attributable to behavioural differences[5]. So for example, assuming that rural and urban households have the same income levels, this income may be a more important factor (implying a bigger coefficient) to rural households as compared to urban ones in their spending decisions.

In order to conduct the Blinder-Oaxaca decomposition as given in equations 7 to 10, the following sample equivalent of the unconditional expectation (equation 4) is employed;

[4]This provides a robustness check of our results to choice of reference group. When decompositions give different conclusions depending on the reference group used, an index number problem is said to obtain. Various attempts have been made in the literature to resolve the index number problem for linear models (e.g. Reimers 1983; Cotton 1988; Neumark 1988; Oaxaca and Ransom 1994).

[5]The coefficient effect in the labour economics literature is interpreted as a measure of discrimination.

$$T\left(\hat{\alpha}_m, \hat{\beta}_m, Z_{im}, X_{im}, \hat{\sigma}_m\right) = N_m^{-1} \sum_{i=1}^{N_m} \left\{ \begin{array}{c} \Phi\left(Z'_{im}\hat{\alpha}_m\right) \Phi\left(\frac{X'_{im}\hat{\beta}_m}{\hat{\sigma}_m}\right) \\ \times \left(X'_{im}\hat{\beta}_m + \frac{\sigma_m \phi\left(\frac{X'_{im}\hat{\beta}_m}{\sigma_m}\right)}{\Phi\left(\frac{X'_{im}\hat{\beta}_m}{\hat{\sigma}_m}\right)} \right) \end{array} \right\} \tag{11}$$

Where $\hat{\alpha}_m, \hat{\beta}_m$, and $\hat{\sigma}_m$ denote sample estimates. With this sample counterpart of the unconditional expectation, equation 7 is estimated by;

$$\begin{aligned} \hat{\Delta}_{R_1}^{DH} = & \left[T\left(\hat{\alpha}_U, \hat{\beta}_U, Z_{iU}, X_{iU}, \hat{\sigma}_U\right) - T\left(\hat{\alpha}_U, \hat{\beta}_U, Z_{iR}, X_{iR}, \hat{\sigma}_R\right)\right] \\ & + \left[T\left(\hat{\alpha}_U, \hat{\beta}_U, Z_{iR}, X_{iR}, \hat{\sigma}_R\right) - T\left(\hat{\alpha}_R, \hat{\beta}_R, Z_{iR}, X_{iR}, \hat{\sigma}_R\right)\right] \end{aligned} \tag{12}$$

Equation 8 is estimated by;

$$\begin{aligned} \hat{\Delta}_{U_1}^{DH} = & \left[T\left(\hat{\alpha}_U, \hat{\beta}_U, Z_{iU}, X_{iU}, \hat{\sigma}_U\right) - T\left(\hat{\alpha}_U, \hat{\beta}_U, Z_{iR}, X_{iR}, \hat{\sigma}_U\right)\right] \\ & + \left[T\left(\hat{\alpha}_U, \hat{\beta}_U, Z_{iR}, X_{iR}, \hat{\sigma}_U\right) - T\left(\hat{\alpha}_R, \hat{\beta}_R, Z_{iR}, X_{iR}, \hat{\sigma}_R\right)\right] \end{aligned} \tag{13}$$

Equation 9 is estimated by;

$$\begin{aligned} \hat{\Delta}_{U_2}^{DH} = & \left[T\left(\hat{\alpha}_R, \hat{\beta}_R, Z_{iU}, X_{iU}, \hat{\sigma}_U\right) - T\left(\hat{\alpha}_R, \hat{\beta}_R, Z_{iR}, X_{iR}, \hat{\sigma}_R\right)\right] \\ & + \left[T\left(\hat{\alpha}_U, \hat{\beta}_U, Z_{iU}, X_{iU}, \hat{\sigma}_U\right) - T\left(\hat{\alpha}_R, \hat{\beta}_R, Z_{iU}, X_{iU}, \hat{\sigma}_U\right)\right] \end{aligned} \tag{14}$$

Finally, equation 10 is estimated by;

$$\begin{aligned} \hat{\Delta}_{R_2}^{DH} = & \left[T\left(\hat{\alpha}_R, \hat{\beta}_R, Z_{iU}, X_{iU}, \hat{\sigma}_R\right) - T\left(\hat{\alpha}_R, \hat{\beta}_R, Z_{iR}, X_{iR}, \hat{\sigma}_R\right)\right] \\ & + \left[T\left(\hat{\alpha}_U, \hat{\beta}_U, Z_{iU}, X_{iU}, \hat{\sigma}_U\right) - T\left(\hat{\alpha}_R, \hat{\beta}_R, Z_{iU}, X_{iU}, \hat{\sigma}_R\right)\right] \end{aligned} \tag{15}$$

If there is only one censoring mechanism, that is $\Phi\left(Z'_{im}\hat{\alpha}_m\right) = 1$, decompositions 7 to 10 reduce to that of a Tobit with censoring from below at zero, as proposed by Bauer and Sinning (2008, 2010) for Tobit models. If expenditure is uncensored at zero, decomposition 7 and 8 are equal, and reduce to the standard Blinder-Oaxaca decomposition with urban coefficients used in the counterfactual. Similarly, decompositions 9 and 10 are equal and reduce to the standard Blinder-Oaxaca decomposition with rural coefficients used in the counterfactual.

3 Empirical Application

The proposed decomposition is used to analyze the rural-urban gap in parental spending on education of own primary school going children in Malawi. Specifically, we want to answer the following question, is the rural-urban spending difference largely due to

differences in characteristics or due to differences in behaviour? The data used in the study come from the Second Malawi Integrated Household Survey (IHS2). This is a nationally representative sample survey designed to provide information on the various aspects of household welfare in Malawi. The survey was conducted by the National Statistical Office from March 2004-April 2005. The survey collects information from a nationally representative sample of 11,280 households. The survey collects annualized household education information which includes household expenditure on primary, secondary, and tertiary education, for household members aged 5 and above. In this illustration, we use husband-wife and single-parent families with at least one child in primary school.

The dependent variable is the share of total annual household expenditure on the education of primary school children in total annual consumption expenditure. In order to account for price variability across areas and time, both expenditure items are deflated by using the Malawi National Statistical Office's spatial and temporal deflator with base national, and February/March 2004. The expenditure items include; fees (tuition and boarding), books and other materials, school uniform, contributions to school building and maintenance, parental association fees, and other school related expenses. We include the following independent variables; age of the youngest primary school going child in the household, the square of age of the youngest child to measure possible nonlinearities, household permanent income as proxied by the log of total household per capita expenditure, proportion of children who go to government schools in a household, number of children in a household, employment status of parents, educational level of parents, parental age, the square of ages for both parents, distance to the nearest primary school to measure quality of access of primary schools.

In order to capture the possibility of gender bias in spending, we construct a variable defined as; $\sum_{i=1}^{10} \frac{H_g}{H}$, where H_g is the number of household members in age-gender group g and H is the household size. We distinguish ten age and gender categories; ages 0-6, 7-15, 16-19, 20-55, and over 55 for each gender. Since we are using aggregate household education expenditure data, this variable can give an indirect test of gender bias in spending. In particular, to check for evidence of differences in spending between primary school going boys and girls we are concerned with the coefficients of the age-gender variable for the ages 7-15 for both sexes. If the coefficients are significant and different that is evidence of preference for a particular sex in spending[6]. We control for regional fixed effects by including a three class regional dummy for the north, centre, and south.

The log of per capita expenditure is potentially endogenous, through two possible channels. Firstly, expenditure and spending on education can be jointly determined through labour supply decisions in the sense that a decision to send children to school

[6]Testing for equality of coefficients in both participation and expenditure equations for all groups of household is done by using a Wald test. This approach to testing for gender bias was first proposed by Deaton (1989).

may be jointly determined with a decision to send the children to work to supplement household income. The second channel for endogeneity would be that parents with a good taste for the education of their children may work harder so they are able to pay for their schooling (Kingdon, 2005). We test for endogeneity using the Rivers and Vuong (1988) procedure for the participation equation, and the Smith and Blundell (1986) procedure for the expenditure equation. Land and its square are our instrumental variables. We find that the log of per capita expenditure is endogenous in the expenditure equation only for rural households. To ensure comparability in terms of number of variables, we included residuals from the reduced form regression for urban households in the urban expenditure equation as well. The reduced form regressions of log of per capita expenditure for both areas, show that the instrumental variables land and its square perform reasonably well as they are significantly correlated with the log of per capita expenditure[7].

Before discussing results of the proposed decomposition, we briefly comment on the DH results for the two areas. For comparison, we also show results of the tobit model (Table 3). Results in Table 1 show that some variables are significant for one group but insignificant for another; an indication of the rural-urban differences. The age of the youngest child is significant and negative only in the participation equation for rural households. This suggests that parents in rural areas are less likely to spend on the education of children as they get older. The level of income as proxied by the log of per capita expenditure significantly increases the likelihood of spending on education and how much is spent for both rural and urban households. The results therefore suggest that income matters at both the extensive and intensive margins for the two groups of households. For rural households, having a higher proportion of children going to government schools significantly increases the probability of spending on them but lowers the share of education expenditure. For urban households having more government scholars lowers the chance of spending on primary education but it has no impact on the share of education expenditure in total expenditure.

We find that the number of children influences positively and significantly the share of education expenditure for rural households, but does not significantly affect the likelihood of spending on education[8]. For urban households having more children increases the likelihood that a household will spend on their education but does not affect the share of expenditure. In terms of parental employment, the results show that for rural and urban households a father's and a mother's employment significantly increases the share of expenditure on education as well as the chance that they will spend on children.

[7]The reduced form regression results are not reported but are available from the author on request.

[8]It is worth recognizing that the number of children is potentially endogenous, if there is a quantity-quality trade off where parents prefer fewer children with a good education. Besides, if there is son preference which affects expenditure on children's education, this may also affect family size. Since we have no valid instruments; we addressed the simultaneity problem arising from the quantity-quality trade by re-estimating the DH models for all groups without number of children; our results largely remained unchanged thus giving us confidence that our results may not be biased due to simultaneity.

With respect to education, we find that the education of both the mother and the father positively and significantly affects the decision whether or not to spend as well as how much to spend on the primary education of their children in both rural and urban areas. The quality of access of primary schools as proxied by distance to the nearest primary school has a negative impact on the participation and the expenditure decisions of both rural and urban households[9].

In terms of the age-gender demographics, the results suggest that having more primary school going boys (i.e. proportion of males aged 7-15) and girls (i.e. proportion of females aged 7-15) significantly and positively impacts on the participation and the expenditure decision levels of rural households. The same is true for urban households. We investigate further to check evidence of gender bias against girls by conducting Wald tests of the equality of the coefficients for proportion of males and females aged 7-15 in the two areas. Results of the tests are shown at the bottom of Table 1. The test results indicate that for rural households there is gender bias against girls at both the participation and expenditure decision levels. For urban households, the Wald test results indicate that there are no statistically significant gender differences at both the intensive and extensive margins. Thus, the Wald tests show evidence of gender bias in favour of boys in rural areas only. Interestingly, we observe that when the tobit model is used (see Table 3), there is no evidence of gender bias in both areas. We present the results and discussion of the decompositions in the next section.

4 Results of the Decomposition

Results of the proposed decomposition are presented in Table 2. For comparison, we also show in Table 4 results of the decomposition for the tobit model. The results indicate that the DH model compared to the tobit model has a lower approximation error, implying that it predicts spending more accurately. The gap in the predicted average share of primary education expenditure between rural and urban households is largely due to differences in characteristics. For example, looking at the expenditure differential when urban coefficients are used in the counterfactual, and we also use the urban variance in the counterfactual, 66% of the gap is due to differences in characteristics of the households, and 34% of the gap is explained by differences in estimated coefficients, hence due to

[9]Distance to the nearest primary school can be endogenous, for example some communities may have a leadership which values education and is more vocal and progressive. This may affect both household schooling decisions as well as placement of schools. Another possible source of endogeneity is that parents with high aspirations for their children may "vote with their feet" by moving to areas where schools are nearer. And this unobserved high aspiration by parents may affect both distance to schooling and schooling decisions. We don't have valid instruments for distance to nearest primary school, so we re-estimated the models without distance to nearest primary school and our results were marginally different from those with distance to nearest primary school thus giving us some level of assurance about the reliability of our results.

behavioural differences. The two aggregate effects are statistically significant at 1%. This result means that if rural and urban household characteristics were to be equalized, 66% of the spending gap would vanish. On the other hand, if the behaviour of rural and urban households was equalized, 34% of the spending gap would disappear. Similarly, when the urban coefficients and the rural variance are used in the counterfactual, the results indicate that the characteristic effect is 67.6% and that 32.4% of the expenditure gap is attributable to differences in coefficients. Both effects are statistically significant. In this case 67.6% (32.4%) of the spending gap would vanish if household characteristics (behaviour) were equalized.

The picture that is emerging from the DH decomposition results is that the gap in spending between rural and urban households largely arises from differences in their characteristics. The same conclusion is arrived at when we ignore the participation equation and use the tobit model (see Table 4). It is however worth noting that decomposition results for the tobit consistently give a higher (lower) measure of the characteristic effect (coefficient effect); which suggests that when we when do not account for the fact that spending is made in two stages, we overestimate (underestimate) the characteristic effect (coefficient effect). In a nutshell, the DH and tobit results suggest that the rural-urban gap in expenditure is mainly due to differences in characteristics; and this finding is robust to choice of both variance and coefficients used in the counterfactual as well as ignoring the participation equation as a censoring mechanism. The robustness of the decomposition results to choice of counterfactual implies that we do not have an index number problem.

5 Concluding Remarks

The paper has proposed an extension of the Blinder-Oaxaca decomposition technique to the independent DH. Using the Second Malawi Integrated Household Survey (IHS2), the paper has applied the proposed decomposition to explain the rural-urban difference in parental spending on own primary school children.

Results from the decomposition show that at least 66% of the expenditure differential arises from differences in characteristics, and about 34% is due to behavioural differences (estimated coefficients) between rural and urban households. This conclusion is robust to choice of coefficients and variance used in the counterfactual. It is also robust to assuming that the zeros in expenditure are entirely a result of a corner solution.

References

Bauer, T. K. and Sinning, M. (2010). Blinder-Oaxaca Decomposition for Tobit Models. *Applied Economics*, 42:1569-1575.

Bauer, T. K. and Sinning, M. (2008). An extension of the Blinder–Oaxaca decomposition to nonlinear models. *AstA Advances in Statistical Analysis*, 92: 197-451

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

Blundell, R. W. and Meghir, C. (1987). Bivariate alternatives to the univariate tobit model. *Journal of Econometrics*, 34:179-200.

Cotton, J. (1988). On the decomposition of wage differentials. *Review of Economics and Statistics*, 70: 236-243.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with applications to the demand for durable goods. *Econometrica*, 39: 829-44.

Deaton, A. (1989). Looking for Boy-Girl Discrimination in Household Expenditure Data. *World Bank Economic Review*, 3: 1-15.

Fairlie, R. W. (1999). The Absence of the African-American Owned Business: An Analysis of the Dynamics of Self-Employment. *Journal of Labour Economics*, 17: 80-108.

Fairlie, R. W. (2005). An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social Measurement*, 30: 305-316.

Green, W.H. (1981). On the asymptotic bias of the ordinary least squares estimator of the tobit model. *Econometrica*, 49: 505-513.

Jones, A. M. (1989). A double-hurdle model of cigarette consumption. *Journal of Applied Econometrics*, 4: 23-39.

Jensen, H. and Yen, S. (1996). Food Expenditures Away From Home by Type of Meal. *Canadian Journal of Agricultural Economics*, 44: 67-80.

Kingdon, G. G. (2005). Where Has All the Bias Gone? Detecting Gender Bias in the Intrahousehold Allocation of Educational Expenditure. *Economic Development and Cultural Change*, 53: 409-451.

Mauldin, T., Mimura, Y. and Lino, M. (2001). Parental Expenditures on Children's Education. *Journal of Family and Economic Issues*, 22: 221-241.

Neumark, D. (1988). Employers' Discriminatory Behaviour and the Estimation of Wage Discrimination. *Journal of Human Resources*, 23: 279-295.

Oaxaca, R. L. (1973). Male-Female Wage Differentials in Urban Labour Markets. *International Economic Review*, 14: 693-709.

Oaxaca, R. L., and Ransom, M. (1994). On discrimination and the decomposition of wage differentials. *Journal of Econometrics*, 61: 5-21.

Rivers, D., and Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39: 347-366.

Smith, R., And Blundell, W. R. (1986). An exogeneity test for a simultaneous equation tobit model with an application to labour supply. *Econometrica*, 54: 679-686.

Smith, M. D. (2003). On dependency in Double-Hurdle models. *Statistical Papers*, 44: 581-595.

Su, S. and Yen S.T. (1996). Microeconometric Models of Infrequently Purchased Goods: An Application to Household Pork Consumption. *Empirical Economics*, 21: 513-533.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26: 24-36.

Table 1: Results of the independent DH by area of residence

Variable	Rural		Urban	
	participation	level	participation	level
Household characteristics				
last child's age	-0.05488***	-0.01060	0.30450	0.00072
	(0.01614)	(0.00883)	(0.23719)	(0.00365)
last child's age^2	0.00104***	0.00022	-0.00767	-0.00009
	(0.00038)	(0.00021)	(0.00710)	(0.00009)
consumption expenditure	0.23207***	0.05227***	0.56821***	0.03576***
	(0.05461)	(0.00355)	(0.01981)	(0.01283)
government scholars	0.76890***	-0.11241**	-2.86960**	0.02045
	(0.10820)	(0.04691)	(1.34249)	(0.01268)
children	0.03128	0.04425***	1.63650**	-0.00016
	(0.04129)	(0.01411)	(0.64050)	(0.00804)
children^2	-0.00132	-0.00132**	-0.14069**	0.00046
	(0.00398)	(0.00056)	(0.07013)	(0.00084)
Parental characteristics				
father works	0.00650***	0.00756***	0.04989***	0.02324***
	(0.00138)	(0.00155)	(0.00224)	(0.00296)
mother works	0.20134***	0.02023***	0.64032***	0.02352***
	(0.05835)	(0.00219)	(0.07506)	(0.00132)
father's education	0.00677***	0.01142***	0.02940***	0.00121***
	(0.00170)	(0.00259)	(0.00354)	(0.00019)
mother's education	0.00683***	0.00865***	0.03234***	0.00231***
	(0.00101)	(0.00148)	(0.00609)	(0.00026)
father's age	0.03908	0.01789	0.90438**	-0.01001
	(0.03091)	(0.01648)	(0.37120)	(0.00628)
father's age^2	-0.00019	-0.00018	-0.00825**	0.00010
	(0.00027)	(0.00015)	(0.00341)	(0.00006)
mother's age	0.04274	0.05428***	-0.51033**	-0.01090
	(0.03055)	(0.01915)	(0.21431)	(0.01311)
mother's age^2	-0.00045	-0.00048***	0.00328*	0.00015
	(0.00028)	(0.00018)	(0.00181)	(0.00014)
School characteristics				
distance primary	-0.00699***	-0.00908***		-0.02440***
	(0.00024)	(0.00084)		(0.00158)
Age-gender composition of household				
males aged 0-6	1.11652**	-0.27137	-8.03960	0.20918*
	(0.55346)	(0.21291)	(5.46235)	(0.11936)
males aged 7-15	1.94601***	0.23238***	6.16139***	0.18465***
	(0.54091)	(0.00950)	(0.09781)	(0.00321)
males aged 16-19	1.05852*	0.30828	-11.41691*	0.26668*
	(0.57515)	(0.22514)	(6.57410)	(0.14466)

males aged 16-19	1.05852*	0.30828	-11.41691*	0.26668*
	(0.57515)	(0.22514)	(6.57410)	(0.14466)
males aged 20-55	0.43034	0.14724	-12.28649**	0.17875*
	(0.50640)	(0.18605)	(5.65072)	(0.10662)
females aged 0-6	0.87586	-0.58748**	7.37600	0.26562**
	(0.54953)	(0.25236)	(5.68817)	(0.12612)
females aged 7-15	1.82512***	0.25020***	7.70956***	0.29012**
	(0.23620)	(0.00930)	(0.40535)	(0.12162)
females aged 16-19	0.33254	0.36888*	-8.92036	0.31863***
	(0.59034)	(0.22344)	(5.63699)	(0.12218)
females aged 20-55	0.63089	0.30406	-3.77753	0.12265
	(0.59596)	(0.23380)	(4.96818)	(0.10029)
females above 55	1.47903**	0.51368*	-4.29883	0.41458**
	(0.71350)	(0.28441)	(6.11054)	(0.18104)
Region				
north	0.17206***	0.13929*	-1.56052*	0.04564
	(0.06414)	(0.07396)	(0.87341)	(0.03108)
centre	0.70344***	0.01791	-0.73286	-0.02001*
	(0.05767)	(0.02882)	(0.61972)	(0.01063)
Controls for endogeniety				
residualcons		-0.19670**		(0.02123)
		(0.08155)		(0.01426)
constant	-5.71966***	-1.95478*	(9.54081)	(0.12453)
	(1.40696)	(1.16150)	(12.48037)	(0.33370)
sigma		0.01358***		0.01182***
		(0.00258)		(0.00160)
Log-likelihood	-6167.27		-2075.03	
P-values of equality of coefficients of males aged 7-15 and females aged 7-15:				
	0.007	0.002	0.52	0.36

Notes: The significance asterisks are defined as follows: * p<0.10, ** p<0.05, *** p<0.01. Numbers in parentheses are standard errors. Residualcons is the residual from the reduced form of log per capita consumption expenditure.

Table 2: Blinder-Oaxaca decomposition of the independent DH

Using the urban variance			
Actual expenditure share gap		0.01	0.01
Predicted expenditure share gap		0.0097***	0.0097***
		(0.0012)	(0.0012)
Characteristic effect		0.0064***	0.0066***
		(0.0011)	(0.0002)
	% of raw gap	66%	68.43%
Coefficient effect		0.0032***	0.0031***
		(0.00041)	(0.00063)
	% of raw gap	34%	31.57%
Counterfactual coefficients		urban	rural
Approximation error		0.0003	0.0003
Using the rural variance			
Actual expenditure share gap		0.01	0.01
Predicted expenditure share gap		0.0097***	0.0097***
		(0.0012)	(0.0012)
Characteristic effect		0.006***	0.0069***
		(0.00057)	(0.0015)
	% of raw gap	0.676	0.7113
Coefficient effect		0.0031***	0.0028***
		(0.0002)	(0.00082)
	% of raw gap	32.40%	28.87%
Counterfactual coefficients		urban	rural
Approximation error		0.0003	0.0003

Notes: The significance asterisks are defined as follows: * p<0.10, ** p<0.05, *** p<0.01. Numbers in parentheses are bootstrapped (1000 replications) standard errors. Approximation error is the difference between the actual expenditure share gap and the predicted expenditure share gap.

Table 3: Results of the Tobit by area of residence

Variable	Rural	Urban
Household characteristics		
last child's age	-0.00044***	0.00191
	(0.00016)	(0.00176)
last child's age2	0.00001**	-0.00008
	(0.00000)	(0.00005)
consumption expenditure	0.00171***	0.01259**
	(0.00018)	(0.00611)
government scholars	0.00337***	0.01023
	(0.00072)	(0.00670)
children	-0.00009	0.00293
	(0.00025)	(0.00351)
children2	0.00006**	0.00009
	(0.00002)	(0.00040)
Parental characteristics		
father works	0.00134***	0.0164***
	(0.00032)	(0.00368)
mother works	0.02277**	0.01213***
	(0.00034)	(0.00129)
father's education	0.00151***	0.01206***
	(0.00004)	(0.00037)
mother's education	0.00431***	0.01074***
	(0.00007)	(0.00049)
father's age	0.00059*	-0.00024
	(0.00030)	(0.00216)
father's age2	-0.00000*	0.00000
	0.00000	(0.00002)
mother's age	0.00077***	-0.00259*
	(0.00023)	(0.00144)
mother's age2	-0.00001***	0.00002*
	(0.00000)	(0.00001)
School characteristics		
distance primary	-0.00042*	-0.01039***
	(0.00022)	(0.00295)
Age-gender composition of household		
males aged 0-6	0.00150	-0.02059
	(0.00346)	(0.03275)
males aged 7-15	0.00871	0.01916
	(0.33700)	(0.23000)

males aged 20-55	0.00291	-0.04197
	(0.00316)	(0.02973)
females aged 0-6	-0.00124	-0.00743
	(0.00345)	(0.03255)
females aged 7-15	0.00867	0.07097
	(0.33400)	(0.97700)
females aged 16-19	0.00640*	-0.00056
	(0.00374)	(0.02995)
females aged 20-55	0.00419	-0.00840
	(0.00378)	(0.03112)
females above 55	0.00874*	0.05879
	(0.00448)	(0.04469)
Region		
north	0.00097	0.01347
	(0.00125)	(0.01054)
centre	0.00266***	-0.00272
	(0.00050)	(0.00388)
residualcons	-0.00246*	-0.00538
	(0.00132)	(0.00683)
constant	-0.04727**	-0.05142
	(0.02112)	(0.11718)
sigma	0.00813***	0.01340***
	(0.00011)	(0.00102)
Log-likelihood	(6211.47)	(2107.53)
P-values of equality of coefficients of males aged 7-15 and females aged 7-15:		
	0.2315	0.5768

Notes: The significance asterisks are defined as follows: * p<0.10, ** p<0.05, *** p<0.01. Numbers in parentheses are standard errors. Residualcons is the residual from the reduced form of log per capita consumption expenditure.

Table 4: Blinder-Oaxaca decomposition of the tobit

Using the urban variance			
Actual expenditure share gap		0.01	0.01
Predicted expenditure share gap		0.0059***	0.0059***
		(0.001)	(0.001)
Characteristic effect		0.0044***	0.0046***
		(0.0004)	(0.0001)
	% of raw gap	74.60%	77.97%
Coefficient effect		0.0015***	0.0013***
		(0.00021)	(0.00041)
	% of raw gap	25.40%	22.03%
Counterfactual coefficients		Urban	rural
Approximation error		0.0041	0.0041
Using the rural variance			
Actual expenditure share gap		0.01	0.01
Predicted expenditure share gap		0.0059***	0.0059***
		(0.001)	(0.001)
Characteristic effect		0.0048***	0.0045***
		(0.00021)	(0.00037)
	% of raw gap	81.56%	76.27%
Coefficient effect		0.0011***	0.0014***
		(0.00026)	(0.00022)
	% of raw gap	18.64%	23.73%
Counterfactual coefficients		urban	rural
Approximation error		0.0041	0.0041

Notes: The significance asterisks are defined as follows: * p<0.10, ** p<0.05, *** p<0.01. Numbers in parentheses are bootstrapped (1000 replications) standard errors. Approximation error is the difference between the actual expenditure share gap and the predicted expenditure share gap.