Jurnal Ekonomi Malaysia 20 (Disember 1989) 61 – 79

Problems in Regression Decomposition of Earnings Differentials

Rahmah bt. Ismail

ABSTRAK

Kertas kerja ini bertujuan melihat masalah yang timbul dalam menggunakan dan menginterpretasikan berbagai-bagai teknik dalam mengira perbezaan perolehan. Didapati penggunaan pembolehubah rambang dalam menganggarkan satu persamaan regresi merupakan teknik yang paling mudah. Bagi teknik ini perbezaan perolehan di antara berbagai-bagai aspek ditunjukkan oleh koefisen pembolehubah rambang. Teknik lain yang boleh digunakan untuk melihat perbezaan perolehan ialah menganggarkan persamaan regresi berasingan bagi kumpulan yang berbeza. Perbezaan perolehan boleh dibahagikan kepada 2 bahagian : disebabkan oleh faktor cirian dan disebabkan oleh perbezaan koefisen. Melalui kaedah ini, didapati penggunaan pembolehubah di kawal tidak menjejaskan pembahagian perbezaan perolehan. Tetapi penggunaan wajaran yang berbeza memberi kesan ke atas pembahagian perbezaan perolehan.

ABSTRACT

This paper examines some problems in implementation and interpretation of earnings differentials decomposition techniques. The utilisation of dummy variable in one single regression seems to be the simplest technique. The coefficient of dummy variable indicates a residual group differences in earnings holding measured characteristics constant. Earnings differences for different groups can be decomposed into two parts by estimating separate regressions: that attributable to differences in group characteristics and that attributable to differences in coefficients. The choice of a reference group for a categorical variable does not affect the estimated differences between parameters. But the decomposition of earnings differences is affected by weights attached to different groups. In estimating log earnings functions, dummy variables are often used for factors such as stratum, region and so on. The

estimated earnings functions are then used to separate the observed earnings differentials, e.g. among different race, into various components. The decomposition of earnings differentials involves weighting which are given to a particular groups. This paper examines some problems in implementation and interpretation of these techniques. The emphasis of this paper is on the technical problems. Empirical results serve to illustrate the arguments. The results indicate that certain aspects of regression decomposition are quite sensitive to model specification and weights chosen, so that they must be interpreted with care.

LOG EARNINGS REGRESSION AND DECOMPOSITION OF EARNINGS DIFFERENCES

Log earnings functions are based on human capital theory developed by Schultz (1960, 1961), Becker (1962, 1964) and Mincer (1970, 1972, 1974,1976). In the analysis of earnings differences the model[1] is extended by adding individual characteristics and dummy variables. Perhaps the simplest technique applied in the analysis of earnings difference is the utilisation of a dummy variable for different race or gender groups in a regression based on data that is pooled for all such groups. The estimation equation is written as,

$$\ln E_i = X_i \beta + Z_i \delta + \mu_i \qquad (1)$$

where E reflects the earnings of individual i, X is a matrix of individual characteristics such as education, age and family background; β is a vector of coefficients of individual characteristics; Z is a dummy variable which equals unity if an individual is a member of a particular group and is zero otherwise and δ is its coefficent. The error term μ_i is assumed to be normally distributed with mean zero and variance σ^2.

The coeficient δ is interpreted as residual group differences in earnings for individuals holding measured characteristics constant. Its value can be positive or negative depending on which group gains an earnings reward. The positive value reflects that the reference group (the category omitted from estimation) has less advantage in earnings as compared to the other group. The intercept of earnings function will then shift up. In contrast, the negative value of δ shows that the reference group has more advantage in earnings, thus, shifts the intercept of earnings function downward.

Table 1 reports the empirical results using specification model (1)[2]. The pooled sampel of full and part-time male workers form the Malaysian Family Life Survey (MFLS) 1976/77 are used in the estimation. Estimates are made for: (a) full-time employees (encompassing earnings from the full-time job only) and (b) all individuals with positive cash and in-kind earnings.[3] The intercept of log earnings profile for Chinese workers is 40.3% higher than for Malays among all full-time employees and 23.7% higher among all employees. The Indian intercept is 7% higher in sample one, but in sample two the intercept is 15.2% lower than the Malay intercept.

TABLE 1. Earnings Equation Estimates by Entire Country

Variable	(1)		(2)	
Intercepts	- 0.964	(- 8.380)	- 0.088	(- 0.550)
CH	0.339	(7.381)	0.213	(3.762)
IN	0.068	(1.845)	- 0.165	(- 3.805)
ED	0.105	(16.645)	0.079	(10.512)
EXP	0.041	(6.071)	0.035	(5.481)
EXP^2	- 0.0005	(- 4.562)	- 0.0005	(- 3.993)
ENGSCH	0.361	(6.535)	0.317	(4.558)
PVTSCH	0.139	(2.705)	0.021	(0.343)
URB	0.039	(2.136)	0.105	(1.978)
WEST	0.038	(0.901)	0.003	(1.162)
AGR	-0.159	(- 3.268)	- 0.169	(- 2.365)
FUL			0.182	(1.898)
SELF			- 0.415	(- 4.239)
R^2	0.3490		0.1158	
N	1778		2704	

Note: t - values in parentheses

The functional form of (1) imposes the restriction that the coefficients β i.e. the implicit returns to characteristics are the same across race and gender groups. As such, estimated group differences in earnings are assumed to be independent of the earnings characteristics. This assumption seems implausible because it is likely that returns to characteristics differ across groups. As such, two portions of decomposition are needed: that attributable to differences in group characteristics and that attributable to differences in coefficients.

A single equation model (1) can be estimated to show both differences above by incorporating interaction terms between race or gender groups and characteristics variable. The model becomes,

$$\text{In } E_i = X_i \beta + Z_i \partial + Z_i X_i \alpha + \mu_i \qquad (2)$$

where α is a vector of coefficients of interaction terms which determine the slope of earnings function. A positive value of α reflects that the slope of earnings function of control group is flatter than the slope of the other group. As α reflects the difference in returns to characteristics by different groups, its positive value also indicates that rate of returns to characteristics is lower for control group.

Using functional form (2) we estimated full-time sample from MFLS by experience cohorts, by occupation and by region. The sample is divided into four experience cohorts: 10 years and under, 11-20, 21-30 and 31 above. Results of the estimation by experience cohorts are in Table 2.[4] The coefficients of the Chinese dummy are positively significant in every cohort of experience. These results indicate that Chinese earn higher wages than Malays in every experience cohort. In the oldest cohort, the Chinese receive 43.6% higher earnings than the Malays; this decreases to 33.3% in the youngest cohort.[5] Indians tend to earn higher wages than Malays in all but the youngest cohort. The coefficents of the interaction terms between schooling and race show that in the two younger cohorts, there are significant differences in returns from schooling between Chinese and Malays. Malays receive higher returns than Chinese.

Occupation form the sample are divided into three groups. The first group is composed of professional, technical and related workers, administrative and managerial workers, sales workers and service workers. The second group consists of agricultural workers and the third group consists of other occupations not included in the first two categories. Table 3 shows the results of the estimation.[6] Earnings gap between Chinese and Malays are present in all occupational groups. The intercept of earnings function for Chinese is higher than that for Malays at 25.6% for the professional group, 44.2% for agricultural, and 30.7% in production and others.[7] There are interacial differences in returns from schooling for the professional group. The Malays gain higher returns then the Chinese and the Indians.

The full-time employee sample is divided into three geographical regions. The northern region includes Kedah, Perlis, Perak and Pulau

TABLE 2. Earnings Equation Estimates: By Experience Cohort

Variable	10 and under	11 - 20	21 - 30	31+
Intercept	- 0.428	- 1.003	- 0.429	- 0.101
	(- 5.882)	(- 7.850)	(- 4.937)	(- 1.173)
CH	0.389	1.078	0.491	0.364
	(3.812)	(6.124)	(3.969)	(3.065)
IN	0.221	0.787	0.183	0.104
	(0.006)	(3.616)	(1.237)	(0.649)
ED	0.169	0.179	0.136	0.078
	(2.015)	(11.682)	(10.140)	(3.611)
CHED	- 0.009	- 0.089	- 0.019	- 0.0007
	(- 5.155)	(- 4.704)	(- 1.054)	(- 0.024)
INED	- 0.004	- 0.066	- 0.010	- 0.008
	(- 0.552)	(- 2.501)	(- 0.513)	(- 0.210)
URB	0.454	0.054	0.052	0.224
	(2.650)	(0.757)	(0.898)	(3.203)
AGR	- 0.473	- 0.024	- 0.227	- 0.183
	(- 1.300)	(- 0.243)	(- 3.116)	(- 2.103)
R^2	0.6375	0.3906	0.376	0.1581
N	141	481	565	591

Note: t - values in parentheses

Pinang. The eastern region consists of Kelantan, Terengganu and Pahang. The western region includes the state of Selangor, Negeri Sembilan, Melaka and Johor. The results of the estimation[8] by occupation are reported in Table 4. Earnings differentials between Chinese and Malays are present in the north and the west, where the Chinese receive 39.5 and 59.9% higher earnings.[9] However, there is no significant Chinese and Malay earnings difference in the east. The negative value of the coefficients of interaction terms in the north and west region between Chinese and year of schooling indicate that the Malays receive higher returns from schooling. In the west region, Malays also receive higher returns to schooling than Indians.

TABLE 3. Earnings Equation Estimates : By Occupation

Variable	(1)	(2)	(3)
Intercept	- 2.046	- 0.101	- 0.869
	(- 7.772)	(- 0.238)	(- 4.342)
CH	1.375	- 0.308	1.396
	(3.514)	(- 0.307)	(4.076)
IN	1.102	- 0.843	0.347
	(2.559)	(- 1.093)	(0.989)
ED	0.186	0.027	0.079
	(14.239)	(0.095)	(5.273)
EXP	0.089	0.018	0.040
	(6.371)	(0.750)	(5.051)
EXP^2	- 0.001	- 0.0004	- 0.0005
	(- 5.515)	(- 1.201)	(- 3.330)
CHED	- 0.052	0.061	- 0.058
	(- 2.586)	(1.262)	(- 3.225)
CHEXP-	0.060	0.012	- 0.044
	(- 2.885)	(0.207)	(- 2.281)
$CHEXP^2$	0.001	0.0001	0.0004
	(2.587)	(0.152)	(1.484)
INED	- 0.085	0.084	- 0.032
	(- 4.371)	(1.744)	(- 1.451)
INEXP	- 0.069	0.046	0.039
	(- 2.492)	(1.229)	(1.483)
$INEXP^2$	0.001	- 0.0007	- 0.0009
	(2.541)	(- 1.363)	(- 1.865)
URB	0.222	- 0.241	0.139
	(3.215)	(- 2.104)	(3.330)
R^2	0.4849	0.2164	0.2213
N	67	288	819

Notes: t - value in parentheses

(1) Professional, technical and related workers; administrative and managerical workers; clerical and related workers; sales and service workers.
(2) Agriculture, animal husbandary and forestry workers; garderners, fishermen and hunters.
(3) Production and related workers, transport equipment operators and laborers and others.

TABLE 4. Earnings Equation Estimates: By Region

Variable	North (1)	East (2)	West (3)
Intercept	- 1.322	- 2.129	- 1.281
	(- 5.721)	(- 6.713)	(- 3.632)
CH	1.629	- 4.544	1.329
	(4.306)	(- 1.421)	(2.726)
IN	0.506	-10.752	0.683
	(1.167)	(- 0.852)	(1.353)
ED	0.120	0.176	0.177
	(9.097)	(9.354)	(9.812)
EXP	0.066	0.081	0.028
	(4.857)	(4.537)	(1.999)
EXP^2	- 0.0009	- 0.001	- 0.0002
	(- 4.485)	(- 3.219)	(- 0.561)
CHED	- 0.047	0.062	- 0.058
	(- 2.623)	(0.918)	(- 2.643)
CHEXP	- 0.072	0.457	- 0.030
	(- 3.207)	(1.866)	(- 1.002)
$CHEXP^2$	0.001	- 0.010	0.0004
	(3.096)	(- 2.127)	(0.692)
INED	- 0.011	1.249	- 0.078
	(- 0.532)	(0.813)	(- 3.346)
INEXP	- 0.025	- 0.801	- 0.009
	(- 0.901)	(- 0.794)	(- 0.308)
$INEXP^2$	0.0004	0.027	0.0001
	(0.871)	(0.802)	(1.172)
AGR	- 0.240	- 0.330	- 0.039
	(- 3.816)	(- 1.722)	(- 0.475)
R^2	0.2867	0.3879	0.4356
N	880	259	639

Notes: t - value in parentheses

(1) North : Kedah, Perlis, Perak, Pulau Pinang

(2) East : Kelantan, Terengganu, Pahang

(3) West : Selangor, Negeri Sembilan, Melaka and Johor

The advantage of this technique is the differences in coefficients and in characteristics which attribute earnings differences can strictly be derived from the estimated coefficients. As such, the value of statistical tests can easily be obtained. Furthermore, from the statistical point of view, estimating a single regression model is better than estimating separate regressions. This is because, an error term which is assumed to be common for every individual is restricted to a single value. But in the analysis of earnings differences, separate regressions are commonly used.[10] Using a method developed by Qaxaca (1973), the following equation can be derived. Suppose, we are comparing the earnings of Malays and Chinese. The average observed log earnings are:

$$\overline{\text{InE}}_m = \Sigma_m \text{ In } E_m/n_m \tag{3}$$

and

$$\overline{\text{InE}}_c = \Sigma_c \text{ In } E_c/n_c \text{ for Chinese} \tag{4}$$

The average observed characteristics are:

$$\bar{X}_m = \Sigma_m \bar{X}_m/_{n_m} \text{ for Malays} \tag{5}$$

and

$$\bar{X}_c = \Sigma_c \bar{X}_c/n_c \text{ for Chinese} \tag{6}$$

where; n_m, n_c = the number of individuals in the observed earnings sample for Malays and Chinese.

Alternatively, the following relationship holds at sample means:

$$\overline{\text{InE}}_m = \bar{X}_m \hat{\beta}_m \tag{7}$$

$$\overline{\text{InE}}_c = \bar{X}_c \hat{\beta}_c \tag{8}$$

Using equation (7) and (8), the gross difference in earnings can be written:

$$\overline{\ln E_c} - \overline{\ln E_m} = \overline{X}_c \hat{\beta}_c - \overline{X}_m \hat{\beta}_m \quad (9)$$

Adding and susbtracting $\overline{X}_m \hat{\beta}_c$ to (9) yields:

$$\overline{\ln E_c} - \overline{\ln E_m} = (\overline{X}_c - \overline{X}_m) \hat{\beta}_c + (\hat{\beta}_c - \hat{\beta}_m)\overline{X}_m \quad (10)$$

Equation (10) decomposes the difference in mean earnings between Chinese and Malays into a term reflecting the difference in characteristics (weighted by Chinese coefficients) plus a term measuring the difference in factor payments (weighted by the Malay characteristics). The later term is interpreted as an unexplained residual that reflects the change in earnings for Chinese if they faced the returns received by Malays. As an unexplained residual, this component reflects a variety of factors including, for example, discrimination, omitted variables and quality of schooling. This technique permits an accounting of the contribution of each characteristics variable to gross earnings differences.

SPECIFICATION OF VARIABLE

The choice of a reference group (the category omitted from estimation) for a categorical (dummy) variable does not affect the estimated difference between parameters within the group, nor does it affect the estimated coefficients for other variables.

Consider the following simple example of regression incorporating a categorical variables ($Z_i = 0, 1; Z_1 + Z_2 = 1$), which a function of categorical variable only. We are interested in measuring earnings differentials between Chinese and Malays. Because Z is dichotomous, equation (7) can be written in two ways,

(i)

$$\overline{\ln E_m} = \alpha_m + Z_{1m}\, \partial_{1m} = (g_m + h_{2m}) + Z_{1m}(h_{1m} - h_{2m}) \quad (11)$$

(ii)

$$\overline{\ln E_c} = \alpha_c + Z_{1c}\, \partial_{1c} = (g_c + h_{2c}) + Z_{1c}(h_{1c} - h_{2c}) \quad (12)$$

or

(i)

$$\overline{\ln E_m} = \alpha_m + Z_{2m}\,\partial_{2m} = (g_m + h_{1m}) + Z_{2m}(h_{2m} - h_{1m}) \qquad (13)$$

(ii)

$$\overline{\ln E_c} = \alpha_c + Z_{2c}\,\partial_{2c} = (g_c + h_{1c}) + Z_{2c}\,(h_{2c} - h_{1c}) \qquad (14)$$

where α is the intercept and ∂ is the coefficient of categorical dummy variable. The coefficient ∂ reflects differences between group (occupation) specific coefficients (h) relative to omitted group. The coefficient for the omitted group is implicitly incorporated into the intercept g.

Based on equation (10) a decomposition of gross earnings differences between race groups can be written in two forms, depending on the choice of control variable.

$$\begin{aligned}
\overline{\ln Ec} - \overline{\ln Em} &= (g_c + h_{2c}) - (g_m + h_{2m}) \\
&\quad + Z_{1c}(h_{1c} - h_{2c}) - Z_{1m}(h_{1m} - h_2) \\
&\quad - Z_{1m}(h_{1c} - h_{2c}) + Z_{1m}(h_{1c} - h_{2c}) \\
&= [g_c + h_{2c}) - (g_m + h_{2m})] \\
&\quad + (Z_{1c} - Z_{1m})\,(h_{1c} - h_{2c}) \\
&\quad + [h_{1c} - h_{2c}) - (h_{1m} - h_{2m})]\,(Z_{1m})
\end{aligned} \qquad (15)$$

or

$$\begin{aligned}
\overline{\ln E_c} - \overline{\ln E_m} &= (g_c + h_{2c}) - (g_m + h_{2m}) \\
&\quad + Z_{2c}(h_{2c} - h_{1c}) - Z_{2m}(h_{2m} - h_{1m}) \\
&\quad - Z_{2m}(h_{2c} - h_{1c}) + Z_{2m}(h_{2c} - h_{1c}) \\
&= (g_c + h_{1c}) - (g_m + h_{1m}) \\
&\quad + (Z_{2c} - Z_{2m})(h_{2c} - h_{1c}) \\
&\quad + [(h_{2c} - h_{1c}) - (h_{2m} - h_{1m})]\,(Z_{2m})
\end{aligned} \qquad (16)$$

The first term of the decomposition reflects differences in intercept between Chinese and Malays.The second term evaluates differences in mean characteristics at the Chinese coefficients and the third evaluates difference in coefficients at Malay means. The second term of equation (15) and (16) are equivalent.[11] But the first and the third terms of each expression are not equivalent. However, their sums are equal. This indicates that the decomposition of group earnings differences into components attribute of to differences in intercepts is not invariant with respect to the choice of a control categorical variable.

The empirical impact of alternative treatment of categorical occupation – specific dummy variables is reported in Table 5. The full-time male sample (sample one) from the MFLS is used in the estimation. The occupational classification is as in Table 3. Table 5 summarizes the decomposition results from regressions of hourly earnings on education, experience, medium of schooling, location and occupation. A separate regressions are run for each racial groups.

The regressions were run thrice by dropping one of the occupational groups each time. The first column shows the contribution of differences in mean characteristics for the relevant group of variable to gross earnings differences. The second column shows the contribution of differences in coefficients. The third column summarizes the variable – specific impacts, the sum of column 1 and 2. Because specification of occupation does not affect other coefficients, the contribution of variables other than occupational group are reported only once.

Consistant with the earlier discussion, column 1 indicates that calculation of the contribution of differences in occupational characteristics to the gross earnings differentials are invariant to the choice of a control group. The second column also show that the sum of the difference in the intercepts and occupational coefficients are equal regardless to the choice of control variable. Setting production workers as a reference category implies that occupational accounts for 1.5% earnings advantage for Chinese. While omission of professional workers from estimation implies that occupation imparts 2.2% earnings advantage for Chinese. We found that experience attributes the largest contribution to earnings differences in favour the Chinese. Furthermore, the partioning of the gross differential between the contribution of difference in characteristics and difference in coefficients (including intercepts) is unaffected by the choice of the control group.

TABLE 5. The Impact of the Choice of a Control Occupational Group Regression Decomposition of Earnings Differences between Chinese and Malays

Variable	Due to Characteristics Differences	Due to Coefficient Differences	Total
ED	0.067	- 0.282	- 0.215
EXP	0.019	- 1.056	- 1.037
EXP^2	- 0.004	0.386	0.382
ENGSCH	- 0.009	- 0.003	- 0.012
ESED	0.054	0.015	0.069
URB	0.018	0.0003	0.0183
Omit Production and Related Workers			
Occupation	0.007	0.008	0.015
Intercept	-	1.271	1.271
Omit Professional Workers			
Occupation	0.007	0.022	0.029
Intercept	-	1.257	1.257
Omit Agriculture Workers			
Occupation	0.007	0.005	0.012
Intercept		-1.274	1.274
Total	0.152	0.339	0.491

ALTERNATIVE WEIGHTS OF REGRESSION DECOMPOSITION

Equation (10) is not unique. The structure of the equation depends on the terms added and substracted to the equation. If the term $\overline{X}_c \hat{\beta}m$ were added and substracted, the following decomposition would hold,

$$\overline{\ln E_c} - \overline{\ln E_m} = (\overline{X}_c - \overline{X}_m)\ \hat{\beta}_m + (\hat{\beta}_c - \hat{\beta}_m)\overline{X}_c \qquad (17)$$

Here, the Chinese weighted of equation (10) is switched to Malay weighted.

Following Reimers (1983) a general earnings differentials between Chinese and Malays can be written as,

$$\overline{\ln E_c} - \overline{\ln E_m} = (\overline{X}_c - \overline{X}_m)[D\hat{\beta}_c + (I-D)\hat{\beta}_m]$$

$$+ [\overline{X}_c(I-D) + \overline{X}_m D](\hat{\beta}_c - \hat{\beta}_m) \quad (18)$$

Where I is the identity matrix and D is a diagonal matrix of weights. The firts term of the right-hand side measures the differences in average characteristics and the second term measures the differences in the parameters of the earnings function that may be caused by labour market discrimination and other ommitted factors. Differences in characteristics are weighted by a linear combination of the coefficients of Chinese and Malays and the coefficient differences are weighted by a linear combination of the characteristics.

The diagonal matrix D can take any possible values depending on what assumption is used. If it is assumed that discrimination penalizes one group (Malays) by preventing its members from earnings accorded to the other group's (Chinese) earning offer function, then D=I. In this case, we assume that Chinese's observed earnings structure would apply to both groups and that is how equation (10) is derived.

On the other hand, we might assume that discrimination gives a preferred group (Chinese) an over-paid wage. In this case D=0 and Malays observed earnings structure would apply to both groups. then, we come up with equation (17). However, since employees preference for one group and their distaste for another group may distort both groups' earnings, neither group's earning offer function would exist in a non-discriminatory world. Instead, the non-discrimination earnings function lies somewhere between them. Therefore, another possibility to observe is D=0.51. The earnings differentials decomposition can be expressed as follows,

$$\overline{\ln E_c} - \overline{\ln E_m} = .5(\overline{X}_c - \overline{X}_m)(\hat{\beta}_c + \hat{\beta}_m) + .5(\overline{X}_c + \overline{X}_m)(\hat{\beta}_c - \hat{\beta}_m) \quad (19)$$

In this case, we are giving equal weight to Chinese and Malay's earnings structure. In other case we might assume D=0.1I to give Malay's observed earnings structure larger weight. As such, a decomposition of earnings differentials is written as,

$$\overline{\ln E_c} - \overline{\ln E_m} = (\overline{X}_c - \overline{X}_m)[.1\,\hat{\beta}_c + .9\,\hat{\beta}_m] + [.9\overline{X}_c + .1\overline{X}_m](\hat{\beta}_c - \hat{\beta}_m) \quad (20)$$

Equation (1) can also be interpreted in the context of regression decomposition. This equation permits different intercepts between groups to differ but restricts factor returns to be the same among Chinese and Malays. Except for the intercept, the contribution of coefficient differences to the gross differential is equal to zero by construction. Thus, the gross earnings differential can be explained in terms of differences in intercepts and characteristics alone.

The quantitative impact of alternative methods of weighting regression decomposition is seen in Table 6. The first panel uses D=0. As such, characteristic differences are weighted by Chinese means. In contrast, panel (2) uses D=I. As such, characteristics differences are weighted by Chinese coefficients and coefficient differences are weighted by Malay means. The third and the fourth panel use D= .1I and D= .5I respectively. The last panel indicates the decomposition using pooled model regression results (equation (1)).

The results is Table 6 indicate that weighthing by D=0 attributes a larger share of gross earnings differentials to the characteristics and a smaller share to difference in coefficients as compared to the results when D=I. Using D=0 attributes about 43.6% to gross earnings differences to the characteristics differences and 31% when D=I. The difference in coefficients attributes 56.4% to gross earnings differences when D=0 and 69% when D=I. The variable specific effects, the sum of the means and coefficient effects for each particular set of variable are invariant with respect to the choice of weights.

The last panel shows that the coefficient on the dummy variable (intercept) is very close to the sum of the coefficient and intercept effect when D=0.5I. This is not an unexpected result as much as the pooled coefficients will reflect a weighted average of estimates derived from separate regression. The restrictive functional form of the pooled model does not appear to affect estimates of the unexplained residual.

CONCLUSIONS

This paper addresses several issues that are frequently overlooked when regression decomposition is used as a tool to analyse earnings differences. In decompositing earnings differentials there are several methods that can be used. The results vary from one method to another. In general,

TABLE 6. Regression Decomposition Results under Alternative Weighting

Weights	Due to Difference in Characteristics $(\bar{X}_c-\bar{X}_m)[D\ \hat{\beta}_c+(I-D)\ \hat{\beta}_m]$	Due to Difference in Coefficients $[\bar{X}_c(I-D)+\bar{X}_m D]\ (\hat{\beta}_c-\hat{\beta}_m)$	Total
1) D=0			
ED	0.109	- 0.324	- 0.215
EXP	0.049	- 1.087	- 1.038
EXP^2	- 0.008	0.391	0.383
ENGSCH	- 0.008	- 0.004	- 0.012
ESED	0.045	0.024	0.069
URB	0.018	0.0005	0.018
AGR	0.009	0.006	0.015
Intercept		1.271	1.271
Total	0.214	0.271	0.491
(Percentage)	(43.6)	(56.4)	(100.0)
2) D=I			
ED	0.067	- 0.282	- 0.215
EXP	0.019	- 1.056	- 1.037
EXP2	- 0.004	0.386	0.382
ENGSCH	- 0.009	- 0.003	- 0.012
ESED	0.054	0.015	0.069
URB	0.018	0.0003	0.0183
AGR	0.007	0.008	0.015
Intercept	-	1.271	-
Total	0.152	0.339	0.491
(Percentage)	(31.0)	(69.0)	(100.0)
3) D= ·1I			
ED	0.105	- 0.320	- 0.215
EXP	0.049	- 1.084	- 1.035
EXP^2	- 0.004	0.390	0.386
ENGSCH	- 0.009	- 0.004	- 0.013
ESED	0.040	0.023	0.063
URB	0.018	0.0005	0.0185
AGR	0.009	0.006	0.015
Intercept	0.208	1.271	1.271
Total	-	0.283	0.491
(Percentage)	(42.4)	(57.6)	(100.0)

continued to next page

TABLE 6. *(Continued)*

4) D= · 51			
ED	0.088	- 0.303	- 0.215
EXP	0.034	- 1.071	- 1.037
EXP^2	- 0.006	0.388	0.382
ENGSCH	- 0.009	- 0.004	- 0.013
ESED	0.050	0.019	0.069
URB	0.018	0.001	0.019
AGR	0.008	0.007	0.015
Intercept	-	1.271	1.271
Total	0.183	0.308	0.491
(Percentage)	(37.3)	(62.7)	(100.0)
5) Pooled Model			
ED	0.107		0.107
EXP	0.050		0.050
EXP^2	- 0.008		- 0.008
ENGSCH	- 0.006		- 0.006
ESED	0.030		- 0.005
URB	0.036		0.036
AGR	0.012		0.012
Intercept	-		0.270
Total	0.221	0.270	0.491
(Percentage)	(37.9)	(62.1)	(100.0)

they imply that the detailed results of regression decomposition experiments must be interpreted with caution. Simple estimation of dummy variables succesffuly provide much of the summary information available from the decomposition methods.

More specifically, the choice of a control group from among a set of categorical variables can affect the decomposition accounting because racial differences in coefficients for the omitted category will be captured in the intercept. Changes in model specification alter the extent to which differences in coefficients and intercepts account for gross earnings differences. However, it does not alter portion of the gross differential accounted for by differences in characteristics.

Finally, analysis of the impact of various weighting algorithms in regression decomposition reveal that simple models utilizing dummy variables in a pooled regression have some favourable properties. These regressions implicitly attach labour force weights to an earnings decomposition which imposes factor neutrality on the unexplained

residual. This specification, however, may not result in the loss of much relevant information due to the sensitivity of the partitioning between the coefficient and intercept effects. The pooled and separate regression give a very similar measures of the unexplained residual.

Appendix 1

Definition of the Variables

In E	Natural logarithm of average of hourly cash earnings plus hourly in-kind earnings, in Malaysian dollars. In kind earnings include food, housing, bonuses, gratuities and others.
ED	Number of years of schooling completed : 0=none; 1-standards 1 (or equivalent), 6=standard 6, 7='remove class' is attended for one year between primary and secondary school by students who change medium of instruction, 8=form one, 12=form five, 13=form six, 14=form upper six, 15=first-year university or college, 22=eight-year university or college.
EXP	Experience defined as age-year of schooling - 6.
EXP^2	Experience square
ENGSCH	Dummy = 1 if medium of instruction of school attended was English.
PVTSCH	Dummy = 1 if school attended was private school
ESED	Interaction between English education and years of schooling completed.
URB	Dummy = 1 if employment was in urban sector.
FUL	Dummy = 1 if individual reports that his employment status in the main occupation is full-time employee.
WEST	Dummy = 1 if individual resides in the West Coast states Johor, Melaka, Negeri Sembilan, Selangor, Perak and Penang.
AGR	Dummy = 1 if the individual's occupation for the job is farmer, farm workers and forestry workers.
SELF	Dummy = 1 if Self-employed workers.
CH	Dummy = 1 if race is Chinese.

IN	Dummy = 1 if race is Indian.
CHED	Interaction between Chinese dummy and years of schooling completed.
CHEXP	Interaction between Chinese dummy and experience.
$CHEXP^2$	Interaction between Chinese dummy and experience squared.
INED	Interaction between Indian dummy and years schooling.
INEXP	Interaction between Indian dummy and experience.
$INEXP^2$	Interaction between Indian dummy and experience squared.

NOTES

[1] The estimated equation of human capital model can be written as:

$$\ln E = \alpha + \beta_1 s + {}_2 t + \beta_3 t^2$$

where E = observed earnings
s = years of schooling
t = years of labour market experience

[2] The estimated equation is:

$$\ln E = \alpha_0 + \alpha_1 CH + \beta_1 ED + \beta_2 EXP + \beta_3 EXP^2 + \beta_4 ENGSCH + \beta_5 PVTS\ CH + \beta_6 URB + \beta_7 WEST \beta_8 AGR + \beta_9 FUL + \beta_{10} SELF + U$$

[3] The full-time employees sample include those who have full-time labour market jobs. Sample two includes full-time employees and self-employed workers. The variables are defined as in Table A.1.

[4] The estimated equation is:

$$\ln E = \alpha_0 + \alpha_1\ CH + \alpha_2\ IN + \beta_1 ED + \beta_2\ CHED + \beta_3\ INED + \beta_4\ URB + \beta_5\ AGR + U$$

[5] These figures are derived at the average of years of schooling as shown below, From the equation in note 4, we derive,

$$\frac{\partial\ \ln E}{\partial\ \ln CH} = \alpha_1 + \beta_2\ \alpha_2\ ED = D$$

Let,

P = percentage difference in earnings then,

$P = e^D - 1$

[6] The estimated equation is:

$$\ln E = \alpha_0 + \alpha_1 CH + \alpha_2 IN + \beta_1 ED + \beta_2 EXP + \beta_3 EXP^2 + \beta_4 CHED + \beta_5 CHEXP + \beta_6 CHEXP^2 + \beta_7 INED + \beta_8 INEXP + \beta_9 INEXP^2 + \beta_{10} URB + U$$

[7] These figures are derived at the average of years of schooling and years of experience and experience square. Method in note 5 is used.

[8] The estimated equation is:

$$\ln E = \alpha_0 + \alpha_1 CH + \alpha_2 IN + \beta_1 ED + \beta_2 EXP + \beta_3 EXP^2 + \beta_4 CHED + \beta_5 CHEXP^2 + \beta_6 CHEXP + \beta_7 INED + \beta_8 INEXP + \beta_9 INEXP^2 + \beta_{10} AGR + U$$

[9] The method used to calculate these figures is as in note 5.

[10] Estimating a single regression to decompose earnings differences involved many interaction terms. To avoid this, estimating separate regressions is preferrable.

[11] This equivalent can be proved by substituting $Z_1 = (1 - Z_2)$ in equation (15).

REFERENCES

Becker, G. S. 1962. Investment in human capital. A theoretical analysis. *Journal of Political Economy* 70: 9-49.

Mincer. J. 1970. The distribution of Labor incomes: A Survey, with special reference to human capital approach. *Journal of Economic Literature* 8: 1–26.

Mincer, J. 1976.. Progress in human capital analysis of the distribution of earnings in *The Personal Distribution of Income*, Edited by A.B. Atkinson, London: Allen by& Unwin.

Oaxaca. R. 1973. Male – Female Wage Differentials in Urban Labor Market. *International Economic Review* 14 (October) : 693-709.

Reimers, C.W. 1983. Labor market discrimination against hispanic and black men. *The Review of Economics and Statistics* 65 (November): 570:579.

Rahmah Ismail. 1987. The effect of human capital on earnings differentials in Malaysia. Ph.D. Dissertation, NCSU, Raleigh.

Rahmah Ismail. 1988. Human capital and earnings differentials in Malaysia: A study by ethnic and activity status. *Penerbitan Tak Berkala* Fakulti Ekonomi, Universiti Kebangsaan Malaysia 36 (Ogos).

Schultz, T. W. 1960. Capital formation by education. *Journal of Political Economy* 68 (December): 57-83.

Schultz, T. W. 1961. Investment in human capital. *American Economic Review* 51 (March): 1-17.

Fakulti Ekonomi
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor, D.E., Malaysia