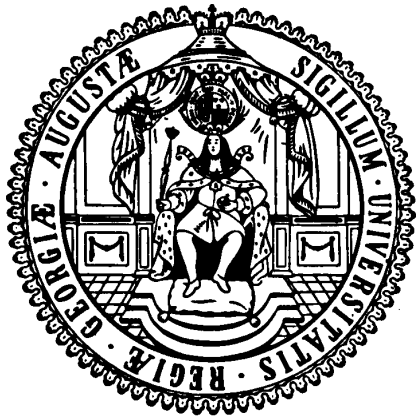
Ibero-Amerika Institut für Wirtschaftsforschung
Instituto Ibero-Americano de Investigaciones Económicas
Ibero-America Institute for Economic Research
(IAI)

Georg-August-Universität Göttingen
(founded in 1737)



Diskussionsbeiträge · Documentos de Trabajo · Discussion Papers

Nr. 131

Inequality and Heterogeneous Returns to Education in Mexico (1992-2002)

Aashish Mehta, Hector J. Villarreal

November 2005

Platz der Göttinger Sieben 3 · 37073 Goettingen · Germany · Phone: +49-(0)551-398172 · Fax: +49-(0)551-398173
e-mail: uwia@gwdg.de · http://www.iai.wiwi.uni-goettingen.de

Inequality and Heterogeneous Returns to Education in Mexico (1992-2002)

by

Aashish Mehta
Asian Development Bank

Hector J. Villarreal*
ITESM Campus Monterrey
Escuela de Graduados en Administración Pública y Políticas Públicas (EGAP)

Within the attempts to understand Mexican economic inequality, returns to education have received a great deal of attention. The driving question has been: why are Mexican wages so unequal? This paper argues that not only the distribution of human capital matters, but also sociodemographic variables, which have their own dynamics and complex interactions with the former. A three-equation maximum likelihood specification in which employment, hours worked and log-wages, as well as their joint variance matrix is proposed, generalizing the Mincerian specification. The resulting is a complex story, where income profiles depend upon particular characteristics.

JEL: O12, J31, D31

* Corresponding Author: hjvp@itesm.mx, CEDES Noveno Piso, Av. Eugenio Garza Sada 2501 Sur, Monterrey N.L. México, C.P. 64849. Tel. (+52-81)8358-2000 ext. 3973.

I. Introduction

Within the attempts to understand Mexican economic inequality, returns to education have received a great deal of attention. The driving question has been: why are Mexican wages so unequal? Is it that the distribution of human capital is very unequal itself, that there is an increasing difference within superior levels of education, or both effects appear combined[1] (Lopez-Acevedo 2004, Meza 1999) ? Moreover, and quite important from a policy perspective: how have the returns to education changed under the time framework? The 90's in Mexico is a period of special interest for the understanding of returns to education (and inequality) because two mayor events occurred. First, signing of commercial treaties and a dramatic increase in trade (e.g. NAFTA). Second, the worst economic crisis in the modern history of the country happened in 1995. A recent paper that tries to understand inequality given these phenomena is Esquivel and Rodriguez-Lopez (2003).

This paper proposes that in order to understand the dynamics of inequality via returns to education, the analysis should not be restricted to the evolution of the educational distribution. There are factors within the workforce population that may be of considerable importance in explaining them. Among these factors different age structures, gender composition, unionization, regional development levels, are instrumental to recognize how returns to education have evolved. In Mexican experience the analysis of these variables has been limited. This should not come as a surprise given that the model employed most of the times to deal with these effects, the Mincerian equation, is quite limited itself (Card 2003).

In this paper we seek a more structural explanation of the returns to education in Mexico, employing a variation of the model first presented in (Mehta and Villarreal 2004). We propose and estimate a three-equation maximum likelihood specification in which employment, hours worked and log-wages, as well as their joint variance matrix, are conditioned on a generalized Mincerian specification. The resulting is a complex story, where the educational levels interact with a set of variables to generate particular

[1] Of course other reasons are pertinent: a quality component within human capital (i.e. different schools' qualities), social networks, etc. However given the availability of data, the two effects mentioned have received most of the attention.

income profiles. If inequality is to be explained via returns to education, these variables need to be considered.

The structure of the paper is the following: Section II presents a description of the empirical environment. Data, variables constructed, a succinct descriptive analysis and some conjectures are included in the section. In Section III the need of a statistical model is motivated, while the model and the function to estimate it are developed in appendices at the end of this study. Econometric results are discussed in this section and simulations of the evolution of income for a specific profile are presented. Section IV links the statistical results with policy implications. Issues regarding development and welfare are considered. Finally, Section V briefly concludes.

II. The empirics

The Mexican education system is a mixture of public and private institutions. The public institutions depend on federal, state or municipal governments for funding. Even though many children attend kindergarten, it is not an official prerequisite for admission to most primary schools.[2] Usually, twelve years of formal education are completed prior to college: six of primary school, three of junior-high and three of high school. College typically takes five years to complete, although the duration does vary.

The data source for this study are the ENIGHs (Encuestas Nacionales de Ingreso y Gasto de los Hogares), which are household income-expenditure surveys, collected by INEGI (Instituto Nacional de Estadistica Geografia e Informatica) in 1992, 1994, 1996, 1998, 2000 and 2002. Our dataset has three strengths. The first is that besides school attainment and income variables, a rich set of sociodemographic characteristics is included in the surveys. Second, it contains data on the successful completion of school years and diplomas, rather than just temporal measures of schooling. As Jaeger and Page (1996) point out, this is important because imputing completion from temporal data can bias results. Thirdly, the surveys have been collected with a consistent methodology, thus enabling intertemporal comparisons.

[2] This may vary according to states or the kind of school. Many private schools do require some preprimary education.

We pare down our sample using criteria that are standard in this literature, restricting our sample to non-students between the ages of 16 (the legal working age) and 65. We include employees and unemployed members of the work-force.[3] Our sample of graduate degree recipients was too thin for computational purposes and we were forced to drop them. Table1 provides the means of most of the variables under consideration in this study.

Description of Variables

Income: refers to quarterly labor income in December 2002 Pesos. The Consumer Price Index (CPI) is employed to put (deflate) the monetary values in equivalent units.

Hours: Is the average number of hours worked per week for the employees within a particular category (sector of the economy[4] or educational level).

Sex Ratio: Is the average of the gender variable within each category. The Dummy variable takes one for male and zero for female, thus a value of 100% implies only males work in that category.

Literate: If the worker is able to read and write. If a worker has an educational attainment of completed primary education, literacy is implied by default.

Union: Is the average of the dummy variable that takes the value of one if the worker belongs to a union, and zero if not.

Age: Average age of the workers within each category.

S_Index: This variable is an index of the average labor income, it uses the year 1992 as base (i.e. 1992=1).

Years: Average years of education for the whole economy and the different sectors.

[3] Smith and Metzger (1998) find, in a Mexican context, that failure to control for returns to capital biases estimates of returns to education upwards as educational attainment correlates positively with capital and earnings. Hence, it is advisable, and standard, to discard observations of self-employed workers.

[4] If an employee works in two different sectors of the economy, his "principal job" is used for classification.

	1992	1994	1996	1998	2000	2002
General						
Income	12291	13601	9326	9890	11814	11671
Hours	44.79	45.24	45.18	45.88	45.62	46.65
sex ratio	71.34%	70.53%	68.72%	68.28%	67.92%	65.83%
Literate	94.30%	93.73%	94.49%	94.34%	95.75%	95.07%
Union	20.36%	16.01%	14.30%	14.70%	15.07%	15.13%
Age	32.13	32.43	32.60	33.22	33.95	34.97
S_index	1.00	1.11	0.76	0.80	0.96	0.95
Years	7.88	8.05	8.34	8.41	8.83	8.79
Primary						
Income labor	8737	8420	6322	6515	7930	7294
Hours	47.18	46.45	46.93	47.01	47.54	46.78
sex ratio	79.05%	77.34%	72.17%	70.63%	71.76%	68.14%
Union	15.47%	12.49%	12.21%	10.47%	9.01%	11.83%
Age	31.19	31.68	32.75	33.36	34.46	35.51
s_index	1.00	0.96	0.72	0.75	0.91	0.83
Secondary						
Income labor	10346	10299	7259	7651	8496	8787
Hours	44.64	44.51	45.13	46.32	46.96	48.28
sex ratio	65.88%	69.08%	71.41%	68.80%	71.69%	68.99%
Union	21.26%	15.63%	12.19%	14.91%	14.06%	13.05%
Age	26.77	27.32	27.63	28.10	29.32	30.55
s_index	1.00	1.00	0.70	0.74	0.82	0.85
High School						
Income labor	15549	16589	10997	12195	12714	13033
Hours	39.86	40.15	42.42	44.01	44.1	46.82
sex ratio	49.50%	45.69%	52.43%	54.00%	51.91%	54.69%
Union	34.64%	27.66%	22.27%	25.46%	23.40%	18.82%
Age	29.75	30.56	30.58	31.32	32.48	32.19
s_index	1.00	1.07	0.71	0.78	0.82	0.84
College						
Income labor	33528	43955	26434	27354	30210	28871
Hours	40.48	43.06	43.33	40.47	40.77	43.93
sex ratio	62.73%	64.43%	59.03%	59.93%	63.24%	60.74%
Union	37.44%	31.25%	29.10%	28.95%	33.07%	27.07%
Age	35.81	35.73	35.47	37.03	37.13	36.61
s_index	1.00	1.31	0.79	0.82	0.90	0.86

Table 1. Aggregated variables for the whole economy and variables by educational level. Mean values of the sample.

Analysis

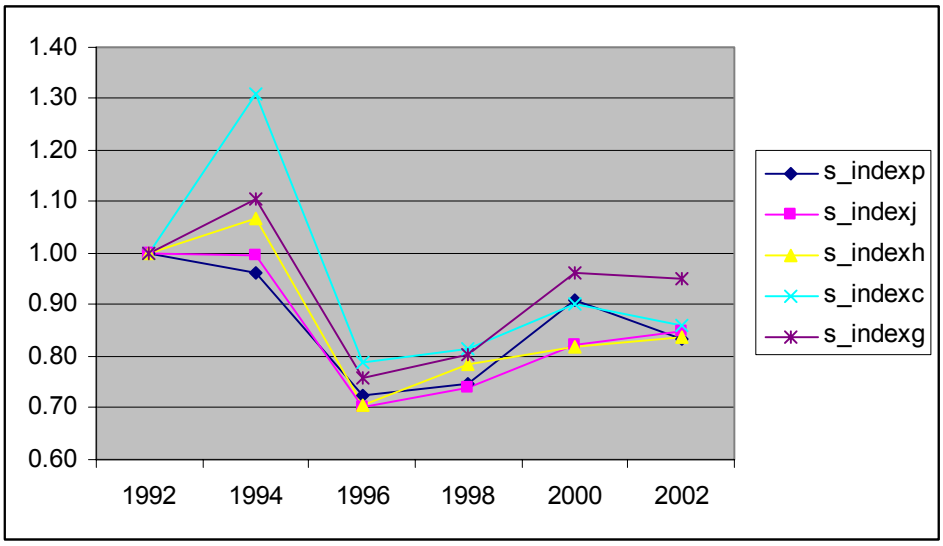


Fig. 1 S_Index for the whole economy and for educational levels.

Figure 1 provides a series of facts for analysis. The first observation is that salaries rose sharply from 1992 to 1994 for the general population. The change from 1992 to 1994 is different when the indices are constructed based on educational levels (completed). In 1994 the labor income of persons with primary education reduces compared to 1992, the labor income of persons with junior high remains close, the earnings of people with high school increase, and finally and quite interesting the earning of people with college overshoot, augmenting close to 30% in real terms. Not surprisingly, studies that that employ 1994 in their data may tend to find an increasing inequality due to education (i.e. Bouillon et. al). Among the explanations provided, a restructuring of the labor markets caused by NAFTA is often invoked. The rationale is that scarce human capital was receiving a premium (Esquivel and Rodriguez-Lopez).

What makes the previous observation starker is that after the 1996 crisis, earnings collapsed in real terms in the general economy and for all the educational levels. The drastic reduction in salaries from 1994 to 1996 has been well documented (Lopez-Acevedo 2004), and despite its magnitude makes sense: the economy had a very strong negative growth, unemployment soared, and prices skyrocketed. It is a well known fact that salaries, due to contracts and other reasons are much stickier compared to most

prices. Consequently they tend to lag with respect to the Consumer Price Index.[5] According to Fig. 1, by year 2002 the general population's average labor income had almost recovered with respect to 1992. The story is different for labor income for the different educational levels: they are at about 85% of those of year 1992. To understand how both facts can be reconciled, the answer may lay on Tables 1: the mean years of education for the general population had increased almost a year, plus people were working on average two more hours per week. Several implications stem from the latter, among which, is that on average people need more human capital and work more hours to obtain similar earnings to those of a decade before. Second, when considering that usually people in the workforce do not receive education, the higher averages imply that persons that are currently joining the workforce have much higher levels of education than their peers, otherwise the increase in average years for the whole sample will occur at a lower pace.

Before presenting any sort of statistical analysis, two caveats should be discussed. The first one, and unfortunately quite common in the literature, is that given the lack of information regarding school quality, a quality component within the human capital, and its evolution cannot be used to explain the observed facts.

The second one and less recognized, is the avoidance of jumping to welfare conclusions from either figures 1 or 2. The reason is that in this paper the CPI is used to deflate, in the presence of very high inflations (in the case of Mexico between 1992 and 2002, it is more than 300%), the consumer price index performs poorly as a cost of living index (Banks et al. 1996, Ruiz-Castillo 1998, Villarreal 2004). Thus, an intertemporal welfare analysis must prudently set its bounds.

[5] Notice that if the contract effect is true, people in the informal sector may have an advantage. If working in the informal sector is correlated with lower educational levels, and these with low incomes; a non intuitive result may be generated: poor people will be less affected by inflation.

III. A statistical framework

If the ultimate goal is the understanding of inequality as related to returns to education, it should be realized that the evolution of returns to education is embedded within a complex dynamic of sociodemographic characteristics. Moreover, if the set of characteristics in the person profile that influences returns to education is big enough, an analysis of means (or differences-in-differences) will not suffice to control or explain the effects. Consequently an econometric analysis needs to be performed.

The workhorse for this kind of analysis is the Mincerian equation, it has nice properties, amongst which it is easy to estimate and interpret. Unfortunately in the presence of heterogeneity, the Mincerian equation tends to predict biased results. Given, the vast variations in the observed dynamics of sociodemographic characteristics for the data in this paper, we need a more structural explanation. We will employ a simplified version of the model presented in (Mehta and Villarreal 2004).

The model utilized in this study and the derivation of the maximum likelihood function used to estimate it, are presented in Appendix A and B respectively. For ease or reading the mathematics and technical details are relegated to the appendices, however the general intuition would be discussed here.

The first part of the job to be realized consists of the estimation of the model for each of the years. Afterwards the implied effects for each sociodemographic effect should be compared across time. In this way the role of sociodemographic characteristics in the income profiles can be identified and their role in inequality inferred. The principal effects will be discussed to some extent in this section. Notice, however that given the large amount of parameters, plus their different economic significance, to sum up the effects and interpret them can be difficult. In order to discuss the results in a more amicable way, the estimated parameters will be employed to simulate and generate a profile that is comparable over time. Of course the patterns do not have to replicate across profiles, and important differences may exist, but the exercise can be done according to the specific group under study.

Results

	92-94	94-96	96-98	98-00	00-02
Union	↑	↑	↑	↓	↑
Experience	↑	↓	↑	↓	↓
Exp. Sq.	↓	↑	↓	↑	↓
Rural	↓	↓	↓	↓	↓
Male	↑	↑	↑	↓	↑
North	↓	↑	↑	↑	↓
South	↓	↑	↓	↑	↓
Primary	↓	↓	↓	↑	↓
Secondary	↑	↑	↓	↓	↑
HighSchool	↑	↑	↓	↑	↓
College	↑	↓	↑	↓	↑

Table 2. Evolution of sociodemographic profile effects on the income equation. An increasing effect (↑) means that the effect was greater that in the previous year, a decreasing effect (↓) means that the effect was smaller that in the previous year.

In the case of North and South, both are dummies referring that the employee lives in that region of the country.[6] The school level effects refer to the effects of years within that level. It is interesting to notice that many of the sociodemographic effects have been increasing or decreasing through time, i.e. there is not a clear tendency. The exception is rural, meaning that ceteris paribus, the effect of living in rural areas is becoming more negative for expected income. One surprising result is the male effect that is the dummy variable of being a male. A decreasing effect was expected, however it is increasing (thus generating a bigger salary gap) for all the comparisons except one.

Experience and experience squared (the obsolesce factor) have not a clear tendency, instead they are flipping signs. This is important because during the next twenty years big cohorts of young workers will incorporate into the labor force. If experience looses importance and new workers come with higher levels of education a displacement effect may occur.

With respect to unionization, as seen in Table 1 there has been some declining in the membership to union in the last ten years, however there positive effect on salaries seems to be increasing in almost all the studied period. To what extent this effect is localized within specific sectors (e.g. government) remains a task for further research.

[6] Twelve states are defined in the North region: BC, BCS, Sin., Son., Chih., Coah., N.L., Tamps., Dgo., Qro., SLP, and Zac. Eight states form the South region: Gro., Oax., Ver., Camp., Chis., Q.R., Tab. and Yuc.

In the case of the parameters employed in this study, almost all of them were statistically significant in explaining the variation of earnings. However, this does not imply that they were economically significant (that is, there effects in earnings may be small), thus further analysis is needed before reaching conclusions. It should also be notice that in this paper, the attention has been put to the effects of parameters on earnings, however and given the nature of the model presented in the appendices, a similar analysis can be made for the propensity of being employed, or the amounts of hours worked.

	1992	1994	1996	1998	2000	2002
0	6771	6629	4712	4918	4712	6574
1	7273	7146	5043	5257	5043	6858
2	7825	7722	5401	5624	5401	7168
3	8433	8367	5788	6021	5788	7505
4	9104	9088	6208	6451	6208	7872
5	9845	9897	6662	6917	6662	8272
6	11733	11420	7878	8451	7878	9740
7	12655	12915	8711	8950	8711	10426
8	13702	14671	9637	9502	9637	11178
9	15731	15354	10596	10984	10596	11772
10	17530	17982	11946	12183	11946	13381
11	19568	21148	13485	13524	13485	15264
12	23011	25033	16084	17290	16084	17728
13	26079	28565	18395	19973	18395	19652
14	29666	32632	21118	23126	21118	21825
15	33875	37321	24336	26840	24336	24286
16	38826	42732	28152	31223	28152	27075
17	44891	55060	32839	35795	32839	31985

Table 3 Simulations of earnings in December 2002 Pesos for a married male, from central Mexico, non-unionized, living in an urban area, with 20 years of experience.

Given the complexity of the effects involved and the amount of parameters, it would be convenient to summarize he complete picture. In order to show how this can be done, some simulations were performed with the parameters estimated with the model of appendices A and B., they are presented in Table 3. A profile that was considered highly representative was chosen[7]: male, married, from central Mexico, non-unionized, living in an urban area, with 20 years of experience. Figure 2 is a good auxiliary to interpret the

[7] Notice the simulations can be performed for each posible sociodemogrpahic profile.

results of Table 3. It seems that for this profile the 1995 crisis had a more or less uniform impact with respect to the 1992 income/education distribution. The impact is not uniform when compared to 1994, because of the overshooting of returns to college years in the 1992-1994 period. Notice that in 1998 are the returns to College years which are showing more improvement, but afterwards they plunged while the lower part of the educational distribution strongly recovers.

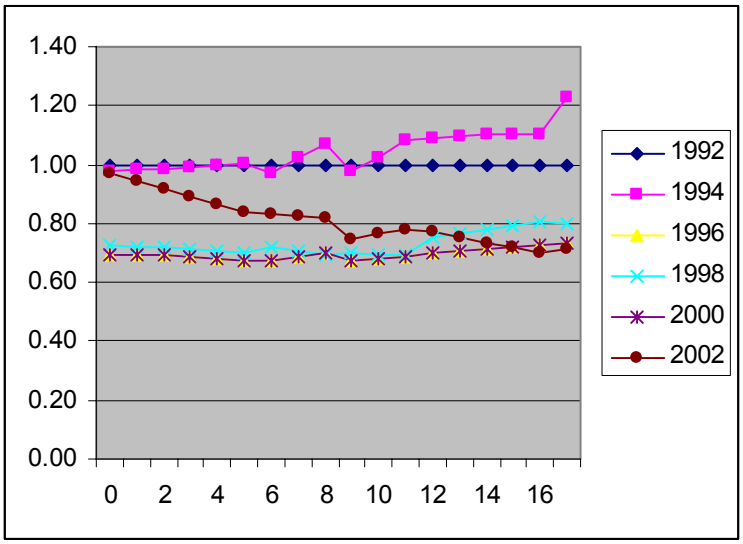


Fig 2.Results from Table 3.

IV. Policy Implications

The effects of additional human capital (i.e. years of education) seem to still have a strong and increasing effect on earnings. However, inequality may persist due to various reasons. First, while the bottom educational distribution have recovered labor income levels similar to those of 1992, the middle and upper part of the distribution are still lagging. This may not be the case for people with graduate school education, but given our sample we cannot measure it.

Second, there exist some specific subgroups that are (or may become vulnerable): people living in rural areas, some people in the south states, older cohorts, and possibly women.

Third, if physical capital is a complement to human physical, it can be conjectured that Mexico may need more physical capital, and that investments of this type, will increase the returns to human capital.

V. Conclusions

This paper started with the purpose to explain in a better way the linkage between returns to education and inequality. It was argued that not only the distribution of human capital matters, but also sociodemographic dynamics that have their own dynamics and complex interactions with the former.

This study suggests that factors such as regional location, age (i.e. experience), gender, rural areas, etc., do play a very significant role in the determination of returns to education and thus inequality. However, the evolution of their effects and their interactions is complex. This invites to perform simulations for more profiles, in order to obtain a detail picture of the educational/income profiles of the whole population. An understanding of inequality and the design of sensible public policy, require it.

Bibliography

Banks, James, Richard Blundell, and Arthur Lewbel. "Tax Reform and Welfare Measurement: Do We Need Demand System Estiamtion", Economic Journal, 106 (1996), pp.1227-1241.

Bouillon, Cesar, Arianna Legovini, and Nora Lustig, "Rising Inequality in Mexico: Household Characterisitcs and Regional Effects", *Journal of Development Studies*, 112-133, (2003).

Card, David, "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems", *Econometrica*, 69, 1127-1160, (2001).

Esquivel, Gerardo, and Jose Antonio Rodriguez-Lopez, "Technology, trade, and wage inequality in Mexico before and after NAFTA", *Journal of Development Economics*, 72, 543-565, (2003).

Goldberger, Arthur, *A Course in Econometrics*, (Cambridge, Massachusetts: Harvard University Press, 1993).

Greene, William, *Econometric Analysis*, (New York: Macmillan Publishing Company, 1990).

Jaeger, David and Marianne Page, "Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education", *Review of Economics and Statistics*, 78, 733-740 (1996).

Lopez-Acevedo, Gladys, "Mexico: Evoluiton of Earnings Inequality and Rates of Returns to Education (1988-2002)", *Estudios Economicos*, 211-284, (2004).

Meza, Liliana, "Cambios en la Estructura Salarial de Mexico en el periodo 1988-1993 y el Aumento en el Rendimiento de la Educacion Superior", *El Trimestre Economico*, 66, 189-226, (1999).

Mehta, Aashish, and Hector J. Villarreal, "What Causes the Sheepskin Effect? An Expanded Mincerian Framework Applied to Mexico", *University of Wisconsin-Madison Working Paper*, (2004).

Ruiz-Castillo, Javier, "A Simplified Model for Social Welfare Analysis: An Application to Spain, 1973-74 to 1980-81", Review of Income and Wealth, 44 (1998), pp.123-141.

Smith, Paula. and Michael Metzger, "The Return to Education: Street Vendors in Mexico", *World Development*, 26(2), 289-296, (1998).

Villarreal, Hector J., "An Intertemporal Comparison of Income and Welfare for Two Mexican Regions", *ITESM Campus Monterrey Working Paper*, (2004).

Appendix A: The Model.

The Mincerian Equation

The most common Mincerian equation takes the following form:

(A1) $$y = \ln w = \delta^0 + \delta^E E + \delta^{E2} E^2 + \sum_{l=p,j,h,c} \left(\delta^{sl} s_l + \delta^{Dl} D_l \right),$$

where w is a person's hourly earnings, sometimes referred to as their implicit wage. E, potential experience, is the maximum length of time they could have been in the labor force given their age and education. l indexes the level of education (primary, junior-high, high-school and college). s_l measures the number of years of education level l completed, and is therefore bounded between zero and the number of school years required to complete that level. D_l indicates whether the lth diploma was received. The growth rate of wages with years of experience and of schooling at level l are $\delta^E + 2\delta^{E2} E$ and δ^{sl} respectively. Similarly, $\exp\left(\delta^{Dl}\right) - 1$ is the percentage wage increase associated

with receipt of diploma *l* over and above that conferred by completion of the final year of the degree. Typically, δ^0 is permitted to vary with personal characteristics. Notice that a specification that "corrects" for such personal characteristics through δ^0 still imposes constant returns to education and experience with respect to these characteristics.

Our Model:

We are interested in the determinants of three variables: employment (z_i = 0 or 1), hours worked if employed (h_i) , and the logarithm of hourly earnings if employed (y_i). In order to investigate these, we specify the following structure based, in principle, on Heckman's (1974) selection scheme. Each person observed in the cross-section is subscripted by *i*.

$$\text{(A2)} \quad \begin{bmatrix} z_i^* \\ h_i^* \\ y_i^* \end{bmatrix} = \begin{bmatrix} \beta x_{zi} \\ \gamma x_{hi} \\ \delta x_{yi} \end{bmatrix} + \begin{bmatrix} \varepsilon_{zi} \\ \varepsilon_{hi} \\ \varepsilon_{yi} \end{bmatrix}; \; \varepsilon_i = \begin{bmatrix} \varepsilon_{zi} \\ \varepsilon_{hi} \\ \varepsilon_{yi} \end{bmatrix} \sim N(0, \Sigma_i); \; \Sigma_i = \begin{bmatrix} 1 & \rho_{1i}\theta_i & \rho_{2i}\sigma_i \\ \rho_{1i}\theta_i & \theta_i^{\,2} & \rho_{3i}\theta_i\sigma_i \\ \rho_{2i}\sigma_i & \rho_{3i}\theta_i\sigma_i & \sigma_i^{\,2} \end{bmatrix};$$

$z_i = 1$ if $z_i^* \geq 0$ and 0 otherwise;

$h_i = h_i^*$ if $z_i^* \geq 0$ and is unreported otherwise;

$y_i = y_i^*$ if $z_i^* \geq 0$ and is undefined otherwise.

Thus, z_i^* is latent employment propensity while h_i^* and y_i^* are the latent hours and logged earnings potentials – observable only if a worker is employed. Σ_i is a positive definite variance matrix for person *i*. θ_i and σ_i are the standard deviations of the "unexplained" components of the hours and logged earnings potentials respectively. Each of the ρ_{ki} is a correlation coefficient between unobservable components.

The allowance for heteroskedasticity is implemented via the Cholesky decomposition such that:

$$\text{(A3)} \quad \varepsilon_i = A_i u_i; \; u_i \sim N(0, I_3); \; A_i = \begin{bmatrix} 1 & 0 & 0 \\ a_{3i} & a_{1i} & 0 \\ a_{4i} & a_{5i} & a_{2i} \end{bmatrix}; \; a_{ji} = \alpha_j^0 + \boldsymbol{\alpha}_j' \mathbf{x}_{ji}. \; j = 1,..,5.;$$

where $\mathbf{x_{ji}}$ are worker characteristics that may condition the variance matrix. From (A3) it follows that:

(A4)

$$\Sigma_i = V(\varepsilon_i) = A_i A_i' = \begin{bmatrix} 1 & a_{3i} & a_{4i} \\ a_{3i} & a_{1i}^2 + a_{3i}^2 & a_{3i}a_{4i} + a_{1i}a_{5i} \\ a_{4i} & a_{3i}a_{4i} + a_{1i}a_{5i} & a_{2i}^2 + a_{4i}^2 + a_{5i}^2 \end{bmatrix}_i = \begin{bmatrix} 1 & \rho_{1i}\theta_i & \rho_{2i}\sigma_i \\ \rho_{1i}\theta_i & \theta_i^2 & \rho_{3i}\theta_i\sigma_i \\ \rho_{2i}\sigma_i & \rho_{3i}\theta_i\sigma_i & \sigma_i^2 \end{bmatrix}.$$

Standard results regarding the log-normal distribution[8] imply the following expressions for the expectation and standard deviation of hourly earnings $\left(w_i^* = \exp(y_i^*)\right)$ for person *i*:

(A5) $E(w_i^*) = \exp(\delta x_{yi} + \sigma_i^2/2)$,

(A6) $S.D.(w_i^*) = \exp(\delta x_{yi})\sqrt{\exp(2\sigma_i^2) - \exp(\sigma_i^2)}$.

This means that in the presence of conditional heteroskedastity in logged earnings (i.e. $\alpha_2, \alpha_4, \alpha_5, \neq 0$), a homoskedastic model is incapable of predicting not only the second, but also the first moment of the earnings distribution, underestimating the expected earnings for persons subject to above average wage variability. It also means that tests on δ do not suffice to test hypotheses regarding average actual (not logged) wages in a heteroskedastic world.

Next, we delineate the content of the main equations and the Cholesky matrix. Two criteria were used in selecting the conditioning variables. First, would their inclusion allow us to estimate parameters crucial to our hypothesis tests? Second, would their exclusion mingle returns to education for different types of people, resulting in erroneous acceptance of the null of no diploma effects?

Logged wages, y_i^*, and employment propensity, z_i^* are conditioned on exactly the components of the RHS of (2), except that a few intercept shifters are added. Each equation is shifted by gender, region, and urban vs. rural location. Union membership condition earnings, but not employment, as there are almost no unemployed union members. Additionally z_i^* is shifted by marital status and the interaction of marital status and gender. Hence we estimate the following conditional expectations functions:

[8] Greene (1990), p. 64.

$$\beta x_{zi} = \beta^0 + \beta^M D_{Male} + \beta^R D_{Rural} + \beta^U D_{Union} +$$
$$+\beta^{North} D_{North} + \beta^{South} D_{South} + \beta^C D_{Couple} + \beta^{CM} D_{Couple} D_{Male} + \beta^E E + \beta^{E2} E^2 \quad \text{(A7)}$$
$$+ \sum_{l=p,j,h} \left(\beta^{sl} s_l + \beta^{Dl} D_l \right) + D_h \left(\beta^{sc}_{PF} s_c + \beta^{DC}_{PF} D_c \right)$$

$$\delta x_{yi} = \delta^0 + \delta^M D_{Male} + \delta^R D_{Rural} + \delta^U D_{Union} +$$
$$+\delta^{North} D_{North} + \delta^{South} D_{South} + \delta^E E + \delta^{E2} E^2 \quad \text{(A8)}$$
$$+ \sum_{l=p,j,h} \left(\delta^{sl} s_l + \delta^{Dl} D_l \right) + D_h \left(\delta^{sc}_{PF} s_c + \delta^{Dc}_{PF} D_c \right)$$

In principle, we could have conditioned hours, h_i^*, on the same Mincerian variables as employment and logged wages. However, as we do not have good reasons to propose the possibility of diploma effects in the hours equation, we include only four slopes - one for each level of schooling, experience and its square, and the same intercept shifters as are included for the employment equation:

$$\gamma x_{hi} = \gamma^0 + \gamma^M D_{Male} + \gamma^R D_{Rural} + \gamma^U D_{Union} + \gamma^{North} D_{North} + \gamma^{South} D_{South} + \gamma^C D_{Couple} \quad \text{(A9)}$$
$$+\gamma^{CM} D_{Couple} D_{Male} + \gamma^E E + \gamma^{E2} E^2$$

It is clear from (4b) that the variables conditioning a_1 and a_2 will most strongly effect θ and σ respectively. Similarly, ρ_1, ρ_2 and ρ_3 can be conditioned through a_3, a_4 and a_5 respectively. There are likely scenarios wherein an urban location and gender would condition all five elements of the variance matrix. We therefore conditioned each Cholesky element on these two characteristics. Similarly ρ_1, ρ_2, ρ_3 and θ are conditioned on the number of years of schooling.[9] Unionization was supposed to effect θ, σ and ρ_3 for obvious reasons. Finally, in keeping with the discussion of section II, σ was conditioned on the same variables as y_i^*, through a_2, in order to capture diploma effects in second moments. Hence, we fill out $\boldsymbol{\alpha}_i$ $(i = 1,..5)$, to specify the following equations:

$$a_{1i} = \alpha_1^0 + \alpha_1^M D_{Male} + \alpha_1^R D_{Rural} + \alpha_1^U D_{Union} + \alpha_1^S Schooling \quad \text{(A10)}$$

$$a_{2i} = \alpha_2^0 + \alpha_2^M D_{Male} + \alpha_2^R D_{Rural} + \alpha_2^U D_{Union} + \alpha_2^E E + \alpha_2^{E2} E^2 + \quad \text{(A11)}$$
$$\sum_{l=p,j,h} \left(\alpha_2^{sl} s_l + \alpha_2^{Dl} D_l \right) + D_h \left(\alpha_{2,PF}^{sc} s_c + \alpha_{2,PF}^{Dc} D_c \right)$$

[9] The key to table 5 describes how the 'schooling' variable was constructed.

(A12) $a_{3i} = \alpha_3^0 + \alpha_3^M D_{Male} + \alpha_3^R D_{Rural} + \alpha_3^S Schooling$

(A13) $a_{4i} = \alpha_4^0 + \alpha_4^M D_{Male} + \alpha_4^R D_{Rural} + \alpha_4^S Schooling$

(A14) $a_{5i} = \alpha_5^0 + \alpha_5^M D_{Male} + \alpha_5^R D_{Rural} + \alpha_5^U D_{Union} + \alpha_5^S Schooling$.

Appendix B: Derivation of the likelihood function.

The sample is divided between those members of the labor force who are employed ($z_i=1$), and those who are not ($z_i=0$). Hence, if *f( )* denotes the distribution of potential hours and log earnings conditional on employment, the log-likelihood function is of the form:

(B1) $LLF = \sum_{z_i=0} \ln\left(\Pr(z_i = 0)\right) + \sum_{z_i=1} \ln\left\{\Pr(z_i = 1) f\left(y_i^*, h_i^* \middle| z_i = 1\right)\right\}.$

We suppress *i* for notational purposes for the rest of the derivation. Let Φ denote the standard normal cumulative distribution function. As usual:

(B2) $\Pr(z = 0) = \Pr(z^* \le 0) = \Pr(\beta x_z + e \le 0) = \Phi(-\beta x_z).$

Further, the joint density of y^*, h^* and *z* in braces in (B1) can be factored differently, and expressed in terms of the latent z^*, rather than *z*:

(B3) $\Pr(z = 1) f\left(y^*, h^* \middle| z = 1\right) = \Pr\left(z = 1 \middle| y^*, h^*\right) g\left(y^*, h^*\right) = \Pr\left(z^* > 0 \middle| y^*, h^*\right) g\left(y^*, h^*\right),$

where *g( )* is the joint density of y_i^* and h_i^* only.

Following Goldberger (1991), pp.196-97, our normality assumptions (B3) imply that $\left(h^*, y^*\right) \sim N\left(\mu_1, \Sigma_{11}\right)$ and $z^*\big|_{h^*,y^*} \sim N\left(\mu_2^*, \Sigma_{22}^*\right)$, where:

(B4) $\mu_1 = \begin{bmatrix} \gamma x_h \\ \delta x_y \end{bmatrix}$, (A4b) $\Sigma_{11} = \begin{bmatrix} \theta^2 & \rho_3\theta\sigma \\ \rho_3\theta\sigma & \sigma^2 \end{bmatrix}$

(B5) $\mu_2^* = \beta x_z + \frac{\rho_1 - \rho_2\rho_3}{\left(1-\rho_3^2\right)}\left(\frac{h^* - \gamma x_h}{\theta}\right) + \frac{\rho_2 - \rho_1\rho_3}{\left(1-\rho_3^2\right)}\left(\frac{y^* - \delta x_y}{\sigma}\right)$ and

(B6) $\Sigma_{22}^* = 1 - \left\{A^2 + 2\rho_3 AC + C^2\right\} / \left(1-\rho_3^2\right)^2;\ A = \left(\rho_1 - \rho_2\rho_3\right); C = \left(\rho_2 - \rho_1\rho_3\right).$

Thus,

(B7) $\Pr\left(z^* > 0 \middle| y^*, h^*\right) = \Phi\left(\mu_2^* / \Sigma_{22}^*\right)$ and

(B8) $g(y^*, h^*)$ is the bivariate normal pdf characterized by (μ_1, Σ_{11}).

Backwards sequential substitution of (B1)-(B6) yield the log likelihood function.