Estimating the Returns to Education using Multiple Regression through a Modified Mincer Equation Specification

Claudynne Kevynne Marchadesch[1], Justin Gabrielle Manay [1] and Jose Lorenzo Mariano[1]
[1] *De La Salle University - Manila*

Abstract: We employed a modified Mincer equation to estimate the returns to education among Filipino workers using multiple regression. Key findings in the final model after doing backward elimination support the existence of a positive relationship between wage and educational attainment, particularly to those who have completed tertiary education. In addition to schooling, years of potential experience, whether the person lives in a rural/urban environment, region of occupation, marital status, sex, type of primary occupation and nature of employment have been found to be significant predictors of wage.

Key Words: Education; Labor; Mincer equation; Multiple Linear Regression; Philippines

1. INTRODUCTION

The inequality in income, particularly for those living in different regions, working in different industries and being of different sexes, is of particular interest for developing countries where despite the fast paced economic growth, the utilization of the labor force is still sub par. As was discussed by Ng (1974), "employment is a means to reduce poverty". Thus, it is imperative to understand the different factors that affect the wages of the labor force of a country, in order to enumerate the ways that the different industries can adjust in order to provide proper wages to those who work

A notable topic in economics is the relationship between wage and educational attainment. (Schady, 2000) Previous studies indicate a positive relationship between the two, many of which are based on the Mincer earnings function, which states that the log of wage is a linear function of the number of years of completed schooling and the number of years of work experience.

In the Philippines, this is taken to be true, as the level of disparity in educational attainment gives rise to an analogous disparity in income levels. Majority of the population, being least educated, are employed in lesser-paid service jobs and the agricultural sector. Those that are more educated, on the other hand, particularly those who have completed tertiary education, are often employed in higher paying professions. (Lou & Terada, 2009) The disparity can be due to the different requirements or specifications needed for different jobs. Thus, the more educated, the more likely it is to get a higher paying job.

This paper uses a modified Mincer model which takes into account other potential factors to wage such as the household size, marital status and others. (Humphreys, n.d)

2. METHODOLOGY

Starting with the basic Mincer model:

$$\ln(W) = \ln(W_0) + \beta_1 S + \beta_2 E + \beta_3 E^2 + \varepsilon$$

where:

W = Wage
W_0 = Baseline Wage
S = Years of schooling/Education in years

E = Years of potential work experience

$\beta_1, \beta_2, \beta_3$ being constants

We modify the above equation by splitting education into different levels (no grade completed, elementary, high school, vocational, college) and adding them into the model as dummy variables, with the "no grade completed" category being the baseline. This is to account for the possibility of non-linear education premiums at every level. Binelli (2008) posits that education premiums are increasingly convex, meaning that education becomes increasingly important at the college level. Thus, we expect the coefficients to increase as we move from elementary to college.

As for the experience terms in the Mincer model, we chose to include only the linear term, since adding the squared term with the linear term would result in multicollinearity. In our model, experience is measured by potential experience, which is computed as the age of a person minus the number of years he/she spent in school.

We also included other variables which could potentially affect wages such as household size, whether the person lives in a rural/urban environment, region of occupation, marital status and sex, among others.

Furthermore, jobs were classified by type (managerial, professional, skilled labor, among others) and by the nature of employment (permanent, temporary, among others) considering different professions offer different compensation.

Having made all these changes, the modified Mincer model to be used is now of the form:

$$\begin{aligned} ln(W_i) = {} & ln(W_o) + \sum_{j=1}^{4} \beta_j S_j + \beta_5 E \\ & + \sum_{j=6}^{9} \beta_j MS_j + \sum_{j=10}^{18} \beta_j JC_j \\ & + \sum_{j=19}^{20} \beta_j NE_j + \sum_{j=21}^{22} \beta_j JC_j + \beta_{23} R \\ & + \beta_{24} F + \varepsilon \end{aligned}$$

where:

W = Wage

Wo = Baseline Wage

S = Years of schooling split into different levels (*Baseline: No Schooling*, Elementary, High School, Vocational, College)

E = Years of potential work experience

MS = Marital Status (*Baseline: Single*, Married, Separated, Annulled, Widowed)

JC = Job Classification (*Baseline: Elementary Occupation*, 1. Managers, 2. Professional, 3. Technicians and associated professionals, 4. Clerical and support, 5. Service and sales, 6. Skilled agricultural/forestry/fishery, 7 Craft and related trade, 8. Plant and machine operators and assemblers, 9. Armed forces)

NE = Nature of Employment (*Baseline: Permanent*, Short-Term, Different Employer)

L = Location/Region (*Baseline: Luzon*, Visayas, Mindanao, Luzon otherwise)

R = Urbanity (*Baseline: Urban*, Rural)

F = Sex (*Baseline: Male*, Female)

To estimate the experience variable, we made the following assumptions:

- We assumed the typical education sequence, as suggested by the dataset (i.e., Nursery, Kindergarten, Preschool, Grades 1-6, Years 1-4).
- The number of respondents with Grade 7 as their highest educational attainment is negligible, so we removed them from the dataset.
- We assumed that all college courses were finished within the span of four years (with the exception of engineering courses, which we assumed to have been finished in five years).
- For vocational courses, we assumed that they had been finished within the span of two years (with the exception of associate engineering courses, which we assume to have been finished in three years). We also removed respondents with Grade 7 as their highest ducational attainment.

We used Python to clean the dataset and SAS to perform the coding requirements of such a large dataset.

3. RESULTS AND DISCUSSION

3.1 Model

Using Backward Elimination, we get the following regression model:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5.44986	0.03065	6714.83761	31621.5	<.0001
elem_grad	0.13106	0.02955	4.17798	19.67	<.0001
hs_grad	0.09874	0.00725	39.34225	185.27	<.0001
voc_grad	0.19031	0.01135	59.74021	281.33	<.0001
col_grad	0.30874	0.00694	420.07027	1978.19	<.0001
exp	0.00079481	0.00021747	2.83654	13.36	0.0003
married	0.10720	0.00557	78.71042	370.66	<.0001
separated	0.06604	0.01487	4.18749	19.72	<.0001
annulled	0.72654	0.18829	3.16167	14.89	0.0001
rural	-0.18089	0.00505	272.60231	1283.74	<.0001
female	-0.23093	0.00532	399.50720	1881.36	<.0001
visayas	-0.14515	0.00644	108.03314	508.75	<.0001
mindanao	-0.22442	0.00595	302.13081	1422.79	<.0001
manager	0.61432	0.01549	333.90139	1572.41	<.0001
professional	0.86768	0.01092	1339.56241	6308.26	<.0001
technician	0.44295	0.01177	300.82219	1416.63	<.0001
clerical	0.42395	0.01034	356.88383	1680.64	<.0001
service	0.10290	0.00749	40.09443	188.81	<.0001
skilled	0.22742	0.09420	1.23769	5.83	0.0158
craft	0.25934	0.00823	210.74168	992.42	<.0001
plant_machine	0.29217	0.01033	169.93559	800.26	<.0001
armed_forces	0.70572	0.03628	80.36237	378.44	<.0001
short_term	-0.09843	0.00588	59.45972	280.01	<.0001
diff_emp	-0.03097	0.01088	1.72089	8.10	0.0044

(Table 3.1)

$$\begin{aligned} logwage = {} & 0.1311(elem\_grad) + 0.0987(hs\_grad) \\ & + 0.1903(voc\_grad) \\ & + 0.3087(col\_grad) + 0.0008(exp) \\ & + 0.1072(married) \\ & + 0.0660(separated) \\ & + 0.7265(annulled) \\ & - 0.1809(rural) - 0.2309(female) \\ & - 0.1452(visayas) \\ & - 0.2244(mindanao) \\ & + 0.6143(manager) \\ & + 0.8677(professional) \\ & + 0.4430(technician) \\ & + 0.4240(clerical) \\ & + 0.1029(service) \\ & + 0.2274(skilled) + 0.2593(craft) \\ & + 0.2922(plant\_machine) \\ & + 0.7057(armed\_forces) \\ & - 0.0984(short\_term) \\ & - 0.0310(diff\_emp) + 5.450 + \varepsilon \end{aligned}$$

- *logwage* (intercept) being the dependent variable, represents the baseline scenario for someone who is a single male with no educational attainment and work experience, living in an urban area in Luzon with a permanent elementary occupation such as : cleaner, helper, manufacturing worker. It will be the basis in which to compare the categorical independent factors with. Summarily, someone with these attributes will earn a wage of exp(5.44986) = 232.73 pesos.

Years of Schooling

- *elem_grad* is the variable that determines the change to logwage when a person finished within elementary. Keeping all other variables constant we calculated logwage in terms of elem_grad by: exp(0.1311) - 1 = 0.1401. This means that compared to the baseline (having no education), the wage of a person who finishes within elementary school increases by 14.01% .
- *hs_grad* is the variable that determines the addition to logwage when a person finished within high school. The table shows that the coefficient of hs_grad is 0.0987. We calculate logwage, keeping all other variables constant by: exp(0.0987) - 1 = 0.1037. This concludes that relative to someone with no formal education, a person who finishes within high school is likely to increase their wage by 10.37%
- *voc_grad* is the variable that determines the addition to logwage when a person has finished a vocational course. Leaving all other variables constant, we can calculate logwage by exp(0.1903) - 1 = 0.2096. This concludes that the person who finishes within college increases their wage by 20.96% compared to those with no formal education.
- *col_grad* is the variable that determines the addition to logwage when a person has graduated from college. The table shows that leaving all other variables constant, logwage is calculated by: exp(0.3087)-1= 0.3616. Therefore compared to a person with no education, a person who finishes within college has a 36.16% increase in wage.

Experience

- *exp* is the potential work experience of a person. Keeping all other variables constant, logwage can be calculated by: exp(0.0007948) - 1 = -0.000794. This shows that for every 1 year of experience a person has, the wage decreases by -0.0794% compared to a person with no experience. This is counterintuitive, but it can be explained by the way we specified experience in our model. Since *exp* is equal to age minus the number of years of schooling, the model suggests that opting out of schooling to "gain experience" leads to lower wages.

  This does not, however, mean that Mincer's equation is misspecified. It simply means that the number of years of potential work experience may not be suitably used as a proxy for the actual number of years of work experience

Marital Status:

- *married*
- *separated*
- *annulled*

  These variables shows how logwage changes according to a person's civil status, leaving all other variables constant. The logwage is calculated as follows: For married people, exp(0.1072) - 1= 0.1132 . For separated people, logwage is exp(0.0660) - 1 = 0.0682 and for annulled people, logwage is exp(0.7265) - 1 = 1.068. This means that compared to single people, married people's wage increases by 11.32% , whereas for separated people their wages increases by 6.82% and for annulled people's wages increases by 107%

Rural

- *rural* is the variable that explains the change in logwage when a person works in a rural area. As per the table, it shows that the rural coefficient is -0.1809. The logwage is calculated by: exp(-0.1809) - 1 = 0.1655. Therefore when a person lives in a rural area their wage is likely to decrease by 16.55% compared to a person who lives in an urban area.

Location/Region

- *visayas* is the variable that explains the change in logwage when a person works in the Visayas region. According to the table, the logwage is calculated by: exp(-0.1452) - 1 = -0.1351. This shows that when a person works in visayas, their wage is likely to decrease by 13.51% compared to a person working in Luzon, leaving all other variables constant.
- *mindanao* is the variable that explains the change in logwage when a person works in the Mindanao region. The value of logwage is calculated by: exp(-0.2244) - 1 = -0.2010. This concludes that when a person works in Mindanao, their wage is likely to decrease by 20.10% compared to those who work in Luzon, leaving all other variables constant.

Sex

- *female* is the variable that determines the wage change based on a person's sex. The logwage is a calculated by: exp(-0.2309) - 1 = -0.2062. This means, if a person is female they are more likely to earn 20.62% less compared to a male, assuming that all other variables are constant.

Job Classification

- *professional* is the variable that determines the change in logwage when a person has a primary occupation as a professional in different fields. An example would be: Health professionals, teaching professionals and Science and Engineering Professionals. Holding all other variables constant, logwage is calculated by: exp(0.8677) - 1 = 1.381. This shows that when a person is a professional, their wages increases by 138.1% compared to a person who has elementary occupation.
- *technician* is the variable that determines the change in logwage when a person has a primary occupation as a technician. Holding

all other variables constant, logwage is calculated by: exp(0.4430) - 1 = 0.5574. This shows that when a person is a technician by profession, their wage is likely to increase by 55.74% compared to a person who has elementary occupation.

- *clerical* is the variable that determines the change in wage when a person has a primary occupation as a clerk. When all other variables are held constant logwage is calculated by: exp(0.4230) - 1 = 0.5265. Therefore, when a person has a clerk job their wage increases by 52.65% compared to people with elementary occupations.
- *service* is the variable that explains the change in wage when a person works in the service industry as a primary occupation. Holding all other variables constant, the logwage is calculated by: exp(0.1029) - 1 = 0.1084. When a person has a job in services, their wages is likely to increase by 10.84% compared to elementary occupation workers.
- *skilled* is the variable that explains the change in wage when a person has a primary occupation in Agriculture, Fishery and Forestry. As per the table, the coefficient is 0.2274. Therefore the logwage is calculated by: exp(0.2274) - 1 = 0.2554. Leaving all other variables constant, a person working a skilled job will have a wage increase of 25.54% compared to a person working an elementary occupation.
- *craft* is the variable that determines the change in wage when a person has an occupation in craft and related trades such as: Handicraft and Printing and Electrical and Electronics trades. Holding all other variables constant, logwage is calculated by: exp(0.2593) - 1 = 0.2961. When a person has an occupation in crafts, their wages are likely to increase by 29.61% compared to a person working in elementary occupation.
- *plant_machine* is the variable that explains the change in wage when a person has a job as a plant and machine operator and assembler. The logwage is calculated by: exp(0.2922) - 1 = 0.3393. Leaving all other variables constant, a person that has a job in plants will have an increase in wage by 33.93% compared to a person working an elementary occupation.
- *armed_forces* is the variable that explains the change in logwage when a person is employed in the armed forces. Leaving all other variables constant, logwage is calculated by: exp(0.7057) - 1 = 1.025. This means that a person with a job in the armed forces has an increase in their wage of 102.5% compared to the wages of a person working in an elementary occupation.

Nature of Employment

- *short_term* is the variable that determines the change in logwage when a person has a short term occupation. Keeping all other variables constant the logwage is calculated by: exp(-0.0984) - 1 = 0.0937. It shows that when a person has short term employment, their wages decreases by 9.37% as compared to a person with permanent employment.
- *diff_emp* is the variable that determines the change in wage when a person changes their employer or moves from one job to another. Logwage is calculated by: exp(-0.0310) - 1 = -0.0305. With all other variables constant, a person with that changes jobs/employers has a decrease in wage by 3.05% compared to a person with permanent employment.

3.2 Testing:

- ANOVA test

The REG Procedure
Model: MODEL1
Dependent Variable: logwage

Number of Observations Read	37149
Number of Observations Used	37149

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	6878.30760	299.05685	1408.32	<.0001
Error	37125	7883.51074	0.21235		
Corrected Total	37148	14762			

(Table 3.2)

Here, the p-value is less than 0.05. Therefore, we reject Ho (all parameters are zero) and we conclude that the data is a good fit for a linear model.

- Adjusted R^2

Root MSE	0.46081	R-Square	0.4660
Dependent Mean	5.82167	Adj R-Sq	0.4656
Coeff Var	7.91551		

(Table 3.3)

The adjusted R^2 value is 0.4656, which concludes that 46.56% of the variation in logwage is explained by the regression model. This value is comparable to those in Luo & Terada (2009) and Schady (2000).

- Autocorrelation

Durbin-Watson D	1.995
Pr < DW	0.2987
Pr > DW	0.7013
Number of Observations	37149
1st Order Autocorrelation	0.003

(Table 3.4)

To test for autocorrelation, we use the Durbin Watson test. For both Pr < DW (test for positive autocorrelation) and Pr > DW (test for negative autocorrelation) we see that the p-value is more than 0.05. Therefore we can conclude that there is no positive nor negative autocorrelation.

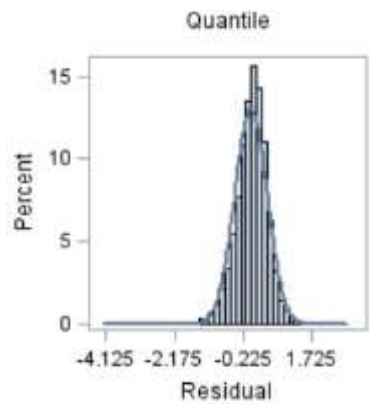
- Normality of errors

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.047712	Pr > D	<0.0100
Cramer-von Mises	W-Sq	31.77168	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	192.0441	Pr > A-Sq	<0.0050

(Table 3.5)

The p-value for the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests are less than 0.05. Therefore we reject Ho (the errors are normal) and assume that the residuals are not normal. However, Ghasemi and Zahediasl (2012) posit that for large samples, normality tests like the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling have high statistical power, meaning that it rejects the null hypothesis with even small deviations from the normal distribution.

From Fig 3.1, the distribution is approximately normal, so it's possible that it may be approximately normally distributed, but the null hypothesis had been rejected due to the large sample size.



(Fig 3.1)

- Homoscedasticity

Bartlett's Test for Homogeneity of residual Variance			
Source	DF	Chi-Square	Pr > ChiSq
group	1	2.3815	0.1228

(Table 3.6)

For homoscedasticity, we conduct Bartlett's test. Because the p-value is greater than 0.05. We do not reject the null hypothesis and conclude that the data is homoscedastic.

- Multicollinearity

Variable	Tolerance	Variance Inflation
Intercept	.	0
elem_grad	0.96418	1.03715
hs_grad	0.68873	1.45195
voc_grad	0.86028	1.16242
col_grad	0.52788	1.89436
exp	0.67535	1.48070
married	0.76551	1.30632
separated	0.92721	1.07850
annulled	0.99842	1.00159
rural	0.90005	1.11104
female	0.86065	1.16191
visayas	0.91343	1.09478
mindanao	0.91162	1.09695
manager	0.85373	1.17132
professional	0.57391	1.74242
technician	0.76191	1.31248
clerical	0.61580	1.62391
service	0.66605	1.50138
skilled	0.99777	1.00224
craft	0.79018	1.26553
plant_machine	0.83730	1.19432
armed_forces	0.97639	1.02419
short_term	0.91136	1.09726
diff_emp	0.90304	1.10737

(Table 3.7)

a. Tolerance

For tolerance to indicate multicollinearity the value must be < 0.1. From Table 3.7, we see that the values for tolerance of all variables are all greater than 0.1. Hence, it can be concluded that there is no multicollinearity.

b. Variance Inflation

The variance inflation factor must be greater than 10 in order to indicate multicollinearity. As seen in the table, the variance inflation factor for all variables are less than 10. Therefore, there is no multicollinearity.

c. Condition Index

For condition index to indicate multicollinearity, the indices must be greater than 30. As per table 3.8, number 24 is the only number with a condition index that is greater 30. However since it is not significantly larger, we can assume that there is no multicollinearity.

Number	Eigenvalue	Condition Index
1	7.09088	1.00000
2	1.46489	2.20012
3	1.12966	2.50540
4	1.05200	2.59623
5	1.04356	2.60671
6	1.02886	2.62525
7	1.01303	2.64569
8	1.00933	2.65053
9	1.00093	2.66163
10	0.99351	2.67156
11	0.98484	2.68329
12	0.96758	2.70711
13	0.92795	2.76431
14	0.91517	2.78355
15	0.75741	3.05974
16	0.59129	3.46299
17	0.52370	3.67968
18	0.46335	3.91198
19	0.39787	4.22165
20	0.22380	5.62889
21	0.18608	6.17298
22	0.16603	6.53515
23	0.06503	10.44214
24	0.00325	46.68900

(Table 3.8)

4. CONCLUSIONS

Having arrived at the adjusted model, there indeed exists a positive relationship between wage and schooling as the parameter estimates for every level of educational attainment is positive. Also, the wage rate increases significantly if one has either graduated from college or from a vocational course. This reinforces the notion that those being the more educated, in particular those who have completed tertiary education are often employed in higher paying professions.

With regards to the primary occupation, professionals are compensated the highest, earning about twice as much as technicians and clerical. Those in the armed forces have the second highest wage followed by managers.

On an interesting note with regards to civil status, annulled workers are the ones who earn most. This is probably because people who pursue annulment tend to be richer. Also, workers who are male, living in urban Luzon, are permanently employed and have not switched their jobs tend to earn more.

There is a disadvantage to those who work in rural areas, Visayas, Mindanao or for those who are female, as they tend to have lower wages.

With regards to the tests, the assumption of normality has not been met; however this can be due to the size of the data.

5. ACKNOWLEDGMENTS

We were encouraged to pursue this research study by our professor Dr. Shirley Ocampo, Mathematics Department, De La Salle University.

6. REFERENCES

- Binelli, C. (2008). Returns to education and increasing wage inequality in Latin America. *Rimini Centre for Economic Analysis Working Paper*, 30-08.

- Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2), 486.

- Luo, X., & Terada, T. (2009). *Education and wage differentials in the Philippines*. The World Bank.

- Ng, L.S (1974). An income and employment consistency model of the Philippines. *Retrospective Theses and Dissertations.* https://lib.dr.iastate.edu/rtd/5388/?utm_source=lib.dr.iastate.edu%2Frtd%2F5388&utm_medium=PDF&utm_campaign=PDFCoverPages

- Schady, N. R. (2000, September 8). What Education Pays?
Non-Linear Returns to Schooling Among Filipino
Men. Retrieved from: https://pdfs.semanticscholar.org/eb9a/37ea20ce3e28eeed638cfbd4e627493e7df7.pdf?fbclid=IwAR2YpXtu_nYLiohY_ImUr0a6Tra13qxB8fm4jnUUqQLwjqkF0cfLpxXmuSE