MINCERIAN (Log-linear) INCOME EQUATION

The mincerian equation of wage determination is the basis of an enormous literature on empirical economics. Jacob Mincer's (1974) wage model is the framework used to estimate returns to education, returns to quality of education, returns to experience, and so on. Mincer developed an income equation that would be dependent on explanatory factors associated with schooling and experience, as well as possibly other attributes, such as gender, for example. Identifying education costs and labor earnings made it possible to calculate the internal rate of return of education, which is the discount rate that equalizes the cost and the expected gain of investing in education.

It is the basis of education economics in developing countries and its estimation has already motivated hundreds of studies, which try to incorporate different educational costs, such as taxes, tuition, opportunity costs, teaching materials, as well as the uncertainty and expectation of the agents at their decisions, technological progress, non-linearity in schooling, etc. It is also used to analyze the relationship between growth and level of schooling of a society, as well as effects on inequality.

One of the great virtues of the Mincerian equation is to incorporate a single equation into two distinct economic concepts:

(a) a price equation revealing how much the labor market is willing to pay for productive attributes such as education and experience and

(b) The rate of return of education, which should be compared with the market interest rate to determine the optimal amount of investment in human capital.

2. Regression Model

The typical econometric regression model derived from the Mincerian equation is:

$$\ln w = \beta_0 + \beta_1 \text{educ} + \beta_2 \exp + \beta_3 \exp^2 + \gamma' x + \epsilon$$

where

w is the wage received by the individual,
educ corresponds to schooling, usually measured by years of study
exp is experience, usually approximated by the age of the individual
x is a vector of individual observable characteristics, such as gender, region. etc
ϵ it's a stochastic error

a. The Coefficient and Attribute Premium

This is a regression model in the log-level format, that is, the dependent variable, the wage is in logarithmic format and the most relevant independent variable, schooling, is in level format. Therefore, the coefficient $\beta 1$ measures how much one year more of schooling causes in proportional variation in the wage of the individual. For example, if $\beta 1$ is estimated at 0.18, this means that each additional year of study is related on average with a wage increase of 18%. This corresponds to the premium of the attribute (or rate of return if the costs were zero). Mathematically, we have:

Deriving, we find that: $(\partial \ln w / \partial \text{educ}) = \beta_1$

On the other hand, by the chain rule, we have:

$(\partial \ln w / \partial \text{educ}) = (\partial w / \partial \text{educ}) (1 / w) = (\partial w / \partial \text{educ}) / w)$

Thus, $\beta_1 = (\partial w / \partial \text{educ}) / w$, corresponds to the percentage variation of the wage from a increase of one year of study..

The coefficient of the mincerian regression with only the constant and a specific variable, say education, gives the gross or uncontrolled relative premium in terms of income variation.

The coefficient of a variable of a multivariate mincerian regression (that is, a log-linear equation with a constant and a series of additional variables) gives us the marginal controlled relative premium in terms of income variation. Thus, a tentative to isolate the effect of this variable from the possible correlations with the other variables considered.

b. The R^2 and Inequality Decomposition[1]

The R^2 of the mincerian regression corresponds to the variance of the log wage explained by the exogenous variables of the regression, that is, the portion of the inequality explained by the set of variables.

The R^2 of a regression with only the constant and a specific variable, say education, gives the gross contribution of that variable to the total inequality (as the Te/T of Theil index).

The R^2 of a regression minus the R^2 of the same regression without one of its variables informs the explanatory power of this omitted variable, controlled by the others, in explaining the total inequality. This corresponds to the marginal contribution of that variable to the total inequality (as the Te/T of Theil index)

---

[1] The coefficient of determination **r^2** (case of two variables) **or R^2** (multiple regression) is a synthetic measure that tells how well the regression line of the sample fits the observed data.

$$R^2 = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2} = \frac{SQE}{SQT}$$

Where:
SQT = SQE + SQR
SQT = Sum of Total Squares (SQT)
SQE = Sum of Explained Squares (SQE)
SQR = Sum of Residual Squares (SQR)

Alternatively, we can define:

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum \left(Y_i - \bar{Y}\right)^2}$$

$$= 1 - \frac{SQR}{SQT}$$

Thus, the R2 quantity defined is known as the determination coefficient (of the sample) and is the measure used as the degree of adjustment of a regression line. Translating, R2 measures the proportion or percentage of the total left-side variation of the equation explained by the regression model.

Applying the Mincerian Equation to Inequality Decomposition Socio-Demographic Determinants

The main results of a multivariate analysis of per capita income distribution between 2001 and 2008 using a mincerian earnings equation approach are presented here. Standard socio-demographic variables such as gender, ethnicity, age, migration status (with respect to states and countries), years of schooling and spatial variables (27 Brazilian States interacting with 5 city sizes (periphery of metropolitan areas, capitals, urban non metropolitan and non capitals and rural areas) are used.

Table below shows the gross contribution of these variables to the inequality of per capita household incomes. This is done simply by means of independent regressions of each variable plus a constant on the particular income concept. The data show first that the six education categories of the household head explain one quarter of total income variation in 2008 - the single most important variable to explain inequality. In 2001, just before the sharp inequality fall, this explanatory power was even higher, 31,31%. A similar effect happens with per capita labor earnings. Geographic variables and ethnicity have also lost explanatory power between 2001 and 2008. Gender kept a constant null explanatory power probably because the results are considered in per capita terms and not at individual levels. Migration increased slightly its explanatory power and so did the age variable, especially when considering all income sources. This capture the effect of the somewhat generous non-contributory pension system turned towards the elderly in Brazil.

Gross Contribution to Income Inequality (in percentages) - **R^2 - CTE + VAR^2**

Per Capita Income		**All Income Sources**		**Labor Earnings**	
	Variable	***2008***	***2001***	***2008***	***2001***
1	***Gender***	***0,0020***	***0,0002***	***0,0305***	***0,0122***
2	***Age***	***8,3227***	***7,0210***	***4,6073***	***4,1649***
3	***Education***	***25,0497***	***31,3089***	***29,0560***	***33,3025***
4	***Ethnicity***	***7,8616***	***10,3042***	***7,0688***	***9,4793***
5	***Migration***	***2,5821***	***2,3392***	***2,0506***	***2,0636***
6	***Geography***	***18,1450***	***21,1074***	***20,6631***	***23,1793***

Inequality decomposition where the interaction between different variables is taken into account is explored in Table below. It basically shows the share of total variation (total R^2) explained when extracting each variable in turn from the complete regression. The data shows similar qualitative results to the previous exercise, namely: a reduction in education explanatory power and an increase in age explanatory power. Geographic variables present an increase in their relative explanatory power.

Table. Net Contribution to Income Inequality – Partial R^2 **(in percentages)**

% Difference of R^2 without a specific Variable with respect to full regression R^2

Per Capita Income		**All Income Sources**		**Labor Earnings**	
	Variable	***2008***	***2001***	***2008***	***2001***
1	***Gender***	***0,2046***	***0,0918***	***0,3178***	***0,1605***
2	***Age***	***14,3245***	***10,2909***	***5,5695***	***3,8033***
3	***Education***	***34,2615***	***35,4399***	***35,7792***	***35,4216***

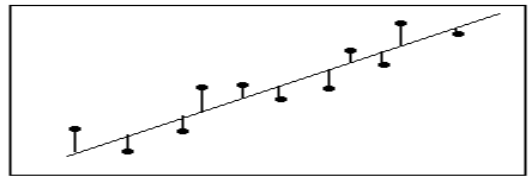
[2] For example in the case of education: $\ln w = \beta_0 + \beta_1 \text{educ} + \epsilon$

What Is Goodness-of-Fit for a Linear Model?

Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals.

In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

Before you look at the statistical measures for goodness-of-fit, you should check the residual plots. Residual plots can reveal unwanted residual patterns that indicate biased results more effectively than numbers. When your residual plots pass muster, you can trust your numerical results and check the goodness-of-fit statistics.



Definition: Residual = Observed value - Fitted value

What Is R-squared?

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

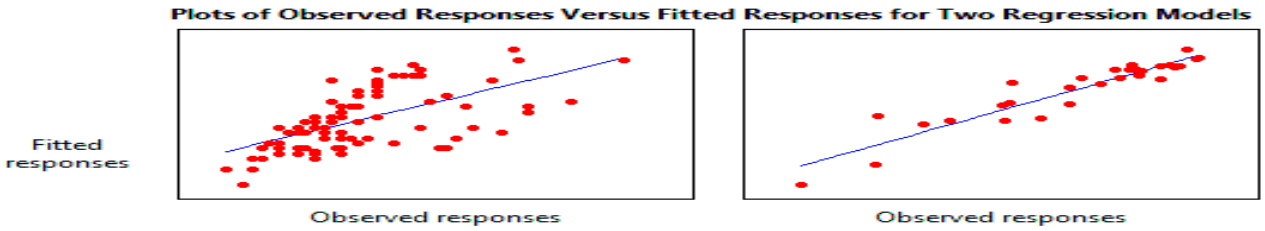
- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

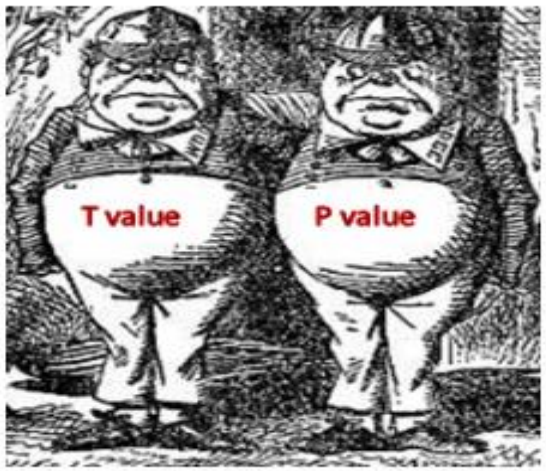
In general, the higher the R-squared, the better the model fits your data. However, there are important conditions for this guideline that I'll talk about both in this post and my next post.

Graphical Representation of R-squared

Plotting fitted values by observed values graphically illustrates different R-squared values for regression models.

Are B'sDifferent from zero?





All in all, the main result to be stressed in these inequality decomposition exercises with respect to socio-economic variable is the reduction of education explanatory power. This may be explained by the increase of the supply of education with the relative stagnation of Brazilian labor markets until 2004. Incidentally, this is the mirror image of seminal work of Carlos Langoni (1973) about inequality increase in Brazil during the 1960s. Even after the reduction of gross explanatory power of years of schooling in income and labor earnings inequality the last seven years it remains by far the most important predictor available for income distribution.

i. Difference in difference estimator

Example of methodology applied to two different periods

In economics, vast research is done analyzing the so-called experiments or quasi-experiments. To analyze a natural experiment it is necessary to have a control group, that is, a group that was not affected by the change, and a treatment group that was directly affected by the event of interest, both with similar characteristics. In order to study the differences between the two groups, pre and post-event data are needed for both groups. Thus, the sample is divided into four groups: the pre-change control group, the post-change control group, the pre-change treatment group, and the post-change treatment group.

The difference between the differences between the two periods for each of the groups is the difference in difference estimator, represented by the following equation:

$g3 = (y2,t – y1,t) – (y2,c – y1,c)$

Where each *y* represents the mean of the studied variable for each year and group, with the subscript number representing the sample period (*1* for before the change and *2* for after the change) and the letter representing the group to which the data belongs (*c* for the control group and *t* for the treatment group). *g3* is the so-called difference in difference estiator. Once the *g3* is obtained, the impact of the natural experiment on the variable to be explained is determined.

In order to study the impacts of local infrastructure policies between two groups, we need data at least two moments in time for both of them. Our sample is thus four fold. The interactive effect between the treatment group dummy (dT=1; dT=0 (control group omitted category)) and the time dummy (d2 =1; d2=0 (initial instant omitted category), which as we will see gives us the difference-in-difference estimator.

Mathematically, we can represent this difference-in-difference estimator (D-D) used from equations in discrete or continuous variables (for example, in the case of logistic regressions or mincerian-type per capita income equations):

$Y = g0 + g1*d2 + g2*dT + (D\text{-}D)*d2*dT + \text{other controls}$

Falling Inequality 2001 to 2009 – Higher income growth for low income groups*:

1) Taking the variable of greatest interest, the difference-difference estimator (D in D), indicates higher income growth for lower-income groups:
   - Region: Northeast x Southeast → (6% when controlled)
   - State - Maranhão x São Paulo → (12% controlled)
   - Rural Area x Metro Region → (16% controlled)
   - Females X Males → (-1% controlled) *exception
   - Blacks X Whites → (4% controlled)
     Browns X Whites → (5% controlled)
   - Construction X other sectors → (3% controlled)
   - Illiterate/0 years x 12 + years → (40% controlled) *41% not controlled*

2) We present below the interactive term of the last controlled model. Full template can be found in the course webpage.

   Example using Schooling variables, Controls and Constant Ommited

Estimated Regression Coefficients			
Parameter	**Estimates**	**t Value**	**Pr > |t|**
CHAVED2 EDUCA03	-1.5661492	-204.86	<.0001
CHAVED2 EDUCA48	-1.4352210	-185.36	<.0001
CHAVED2 EDUCA812	-1.0193033	-133.97	<.0001
CHAVED2 LIXOEDUCA	-1.4114157	-66.63	<.0001
CHAVED2 ZZZZZEDUCA12	0.0000000	.	.
ANO 2009	-0.1155097	-13.78	<.0001
ANO z2001	0.0000000	.	.
CHAVED2*ANO EDUCA03 2009	0.4021384	41.58	<.0001
CHAVED2*ANO EDUCA03 z2001	0.0000000	.	.
CHAVED2*ANO EDUCA48 2009	0.2548667	25.88	<.0001
CHAVED2*ANO EDUCA48 z2001	0.0000000	.	.
CHAVED2*ANO EDUCA812 2009	0.1474745	15.55	<.0001
CHAVED2*ANO EDUCA812 z2001	0.0000000	.	.
CHAVED2*ANO LIXOEDUCA 2009	0.4048677	14.28	<.0001
CHAVED2*ANO LIXOEDUCA z2001	0.0000000	.	.
CHAVED2*ANO ZZZZZEDUCA12 2009	0.0000000	.	.
CHAVED2*ANO ZZZZZEDUCA12 z2001	0.0000000	.	.