Machine Learning: Lab 1

Jingpeng Hong

February 1, 2022

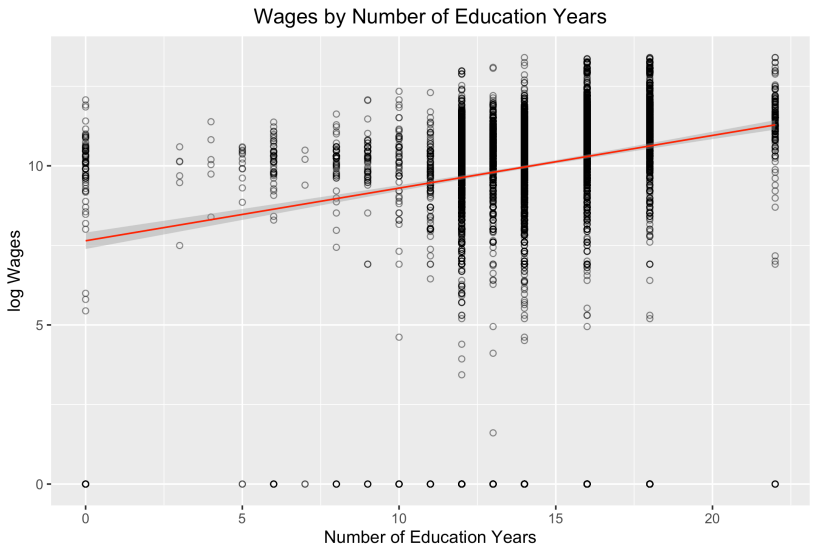
1 Descriptive Statistics

Table 1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
year	8,448	2,020.00	0.00	2,020	2,020
incwage	8,448	56,459.93	69,309.55	0	661,000
lnincwage	8,448	9.99	2.50	0.00	13.40
educdc	8,448	14.20	2.95	0	22
female	8,448	0.48	0.50	0	1
age	8,448	41.90	13.26	18	65
age^2	8,448	1,931.56	1,115.62	324	4,225
white	8,448	0.68	0.46	0	1
black	8,448	0.08	0.27	0	1
hispanic	8,448	0.15	0.36	0	1
married	8,448	0.57	0.49	0	1
nchild	8,448	0.81	1.12	0	9
vet	8,448	0.05	0.21	0	1
hsdip	8,448	0.54	0.50	0	1
coldip	8,448	0.39	0.49	0	1
hsdip#educdc	8,448	7.04	6.51	0	14
coldip#educdc	8,448	6.59	8.31	0	22

2 Scatter Plot

Figure 1: Scatter plot ln(incwage) and education (w/ linear fit line)



3 Baseline Model Estimation

Estimate the following model:

$$\begin{aligned} ln(incwage_i) = \beta_0 + \beta_1 educdc + \beta_2 female + \beta_3 age + \beta_4 age^2 + \beta_5 white \\ + \beta_6 black + \beta_7 hispanic + \beta_8 married + \beta_{10} nchild + \beta_{11} vet + \epsilon_i \end{aligned} \tag{1}$$

3.a

As shown in the first column of Table (??), $R^2 = 0.052$:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{VAR}_{\text{explained}}}{\text{VAR}_{\text{total}}} \tag{2}$$

Hence, 5.2% variation in log wages can be explained by the model.

3.b

Hypothesis Testing:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{10} = 0$$
$$H_A : \beta_j = 0 \quad \text{for some } j$$

As shown in the first column of Table (??), F-statistics $= 45.969$, which is significant at 1% level. Hence, we should reject the null hypothesis H_0 with $\alpha = 0.10$

3.c

The return to an additional year of education:

With an additional year of education, $ceteris\ paribus$, the wage will increase by 15.4% on average.

It is statistically significant, because $t = 0.154/0.009 = 17.11$, where the p-value is less than 0.01. We can reject the null hypothesis that there is no significant difference between the return to education and 0 at the 1% significance level.

It is also practically significant, as the interpretation of the coefficient, one additional year of education will lead to 15.4% increase in the wage. The magnitude of the effect is quite big.

3.d

$Ceteris\ paribus$, the function of the age will be:

$$ln(incwage_i) = 0.115 age_i - 0.001 age_i^2 + constant \tag{3}$$

The model will predict an individual to achieve the highest wage at $0.115/2 \times 0.001 = 57.5$ years old.

3.e

The model will predict that, *ceteris paribus*, men will have higher wages because the coefficient of $female$ (-0.310***) is significantly negative. First, we don't take the industry into account in our model. For example, men will systematically work in the industries with higher wages. Second, women are likely to have less years of experience and less worked hours due to the family leave or other issues. In this model, we didn't control for the working years and hours.

Table 2: OLS Regression

	Dependent variable:	
	lnincwage	
	(1)	(2)
educdc	0.154***	0.059***
	(0.009)	(0.022)
female	−0.310***	−0.315***
	(0.054)	(0.054)
age	0.115***	0.107***
	(0.016)	(0.016)
age^2	−0.001***	−0.001***
	(0.0002)	(0.0002)
white	−0.010	−0.001
	(0.075)	(0.075)
black	−0.062	−0.029
	(0.117)	(0.117)
hispanic	−0.167*	−0.152*
	(0.089)	(0.089)
married	0.138**	0.128**
	(0.063)	(0.063)
nchild	−0.080***	−0.073***
	(0.027)	(0.027)
vet	0.050	0.068
	(0.130)	(0.130)
hsdip#educdc		0.044***
		(0.012)
coldip#educdc		0.063***
		(0.014)
Constant	5.672***	6.443***
	(0.322)	(0.365)
Observations	8,448	8,448
R^2	0.052	0.054
Adjusted R^2	0.051	0.053
F Statistic	45.969*** (df = 10; 8437)	40.410*** (df = 12; 8435)
Note:		*p<0.1; **p<0.05; ***p<0.01

3.f

White: Compared with other non-black races (like Asian, American Indian or Alaska Native), *ceteris paribus*, those white individuals will have 1.0% lower wages, which is not statistically significant.

Black: Compared with other non-white races (like Asian, American Indian or Alaska Native), *ceteris paribus*, those black individuals will have 6.2% lower wages, which is not statistically significant.

Hispanic: Compared with non Hispanic origin, *ceteris paribus*, those Hispanic individuals will have 16.7% lower wages, which is significant at the 10% level.

3.g

Hypothesis Testing:

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$$

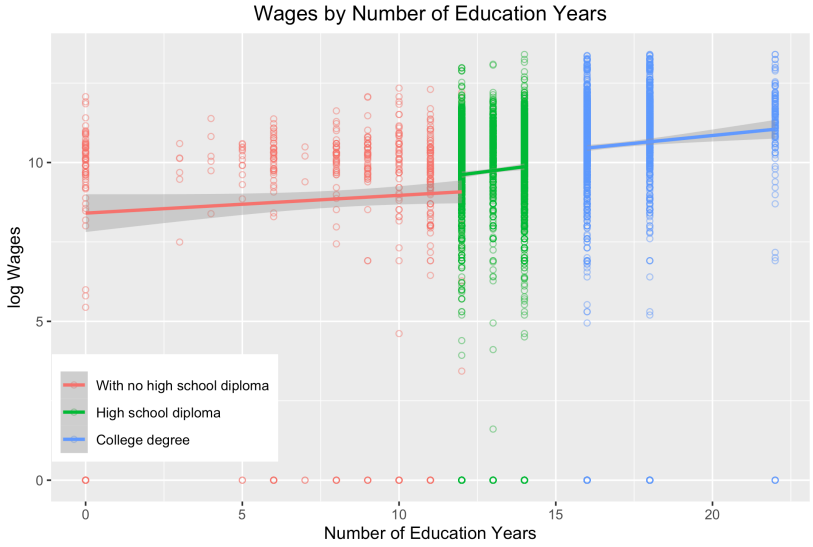
$$H_A : \beta_5 \neq 0 \text{ or } \beta_6 \neq 0 \text{ or } \beta_7 \neq 0$$

The null hypothesis is that race/ethnicity has no effect on wages, which means the coefficients of white, black and hispanic have no significant difference with 0.

The calculation is shown in the R, and the F-statistics of this joint test is 1.504, with p-value $= 0.211$. Hence, we cannot reject the null hypothesis, *i.e.*, we find no evidence that race/ethnicity has significant effect on wages.

4 Scatter Plot by Diploma

Figure 2: Scatter plot ln(incwage) and education by diploma (w/ linear fit line)



5 "Sheepskin" Effects

The model for estimation:

$$\begin{aligned} ln(incwage_i) = \beta_0 &+ \beta_1 educdc + \beta_2 female + \beta_3 age + \beta_4 age^2 + \beta_5 white \\ &+ \beta_6 black + \beta_7 hispanic + \beta_8 married + \beta_{10} nchild + \beta_{11} vet \\ &+ \beta_{12} hsdip \times educdc + \beta_{13} coldip \times educdc + \epsilon_i \end{aligned} \tag{4}$$

In this model, for those with no high school diploma, the return to an additional year of education is β_1; for those with high school diploma, the return to an additional year of education is $\beta_1 + \beta_{12}$; for those with college diploma, the return to an additional year of education is $\beta_1 + \beta_{13}$. The coefficients β_{12} and β_{13} represent the "sheepskin" effect. For individuals with a high school diploma, they will expect to have a higher return of an additional year of education by β_{12} than those without a high school diploma; for individuals with a college diploma, they will expect to have a higher return of an additional year of education by β_{13} than those without a high school diploma

6 "Sheepskin" Effects: Model Estimation

The results estimated are shown in the Column 2 of Table (2)

6.a

The predicted ln(incwage) of the individual with a high school diploma is 9.133418, $i.e.$ incwage=9,259.6

The predicted ln(incwage) of the individual with a high school diploma is 9.843719, $i.e.$ incwage=18,839.7

6.b

Yes, individuals with college degrees will have higher predicted wages than those without, $ceteris\ paribus$.

First, those with college degrees will generally have 4 years more of education, which leads to a 5.9%×4=23.6% higher wages. This is the effect of additional years of education.

Second, those with college degrees will have 6.3%-4.4%=1.9% higher returns for each additional year of education. For 12 years of education, there will be 12×1.9%=22.8% higher wages. This is the "Sheepskin" effect.

Third, those additional 4 years will have an additional effect of 4×6.3%=25.2%.

Therefore, individuals with college degrees will have higher predicted wages than those without by 71.6%

6.c

Given our results in 6.b, college degrees will increase wages significantly both statistically and practically. Hence, legislation expanding access to college education will be beneficial based on the cost-benefit analysis.

7 Discussion

First, we can add the experience into our model. The experience can be simply calculated by Experience = Age - Education. The introduction of the Experience (and Experience^2) will help us reduce the potential omitted variable bias. Since an additional year of education will lead to one year less in experience, we may overestimate the returns of education by missing the return of working experience.

Second, in the IMPUS data we can have the information of Occupation/Industry. We can take the Occupation/Industry fixed effect in our model. Since individuals with different education levels will choose different occupations, leading to the omitted variables bias.

Third, there is potential sample selection bias in this model. For those with lower expected wages, they are likely to be unemployed and cannot be observed in our sample. Hence, we can technically choose models like Heckman two-steps to deal with the sample selection bias.