教育回报的实证分析

基于基本线性回归模型

刘晶芳 (15320171151900)　余星月 (15320171151888)

数据来源：CHFS2013 数据库

研究问题

教育对工资的影响

变量设置

被解释变量:

wage：小时工资；lwage：小时工资对数

核心解释变量:

edu：教育程度；female：性别(1=男性，0=女性)；femedu：性别*教育程度

控制变量 X:

kidsum：7-18 岁孩子个数；childrensum：0-6 岁孩子个数；

health：健康状况；age：年龄；exper：工作年数；

exper2：工作年数平方；spousewage：配偶工资；

spouseedu：配偶教育程度

数据描述

VARIABLES	N	mean	sd	min	max	Var	p50
age	2,902	39.87	7.651	19	60	58.54	40
edu	2,902	12.95	3.589	0	22	12.88	15
kidsum	2,902	1.019	0.573	0	4	0.328	1
childrensum	2,902	1.018	0.571	0	4	0.326	1
married	2,902	0.987	0.112	0	1	0.0126	1
female	2,902	0.501	0.5	0	1	0.25	1
health	2,902	0.592	0.492	0	1	0.242	1
laborforce	2,902	1	0	1	1	0	1
exper	2,902	11.48	9.125	0.1	41	83.27	9
hours	2,902	517.8	182.4	0.6	2,016	33,286	480

wage	2,902	87.93	191.3	0	6,667	36,612	60
exper2	2,902	215.1	283.7	0.01	1,681	80,470	81
spouseedu	2,902	12.78	3.587	0	22	12.86	12
spousehours	2,902	544.2	189.2	1	2,016	35,796	480
spousewage	2,902	140.3	2,746	0	146,732	7.54E+06	58.33
lwage	2,895	4.075	0.834	0	8.805	0.696	4.094

本次数据的样本量为 2902，其中 female 的均值为 0.501，即样本中男女比例相等。教育水平的均值为 13 年，中位数为 15 年，教育水平大体集中在高中、大学水平。而配偶的教育水平平均值接近 12 年，且中位数为 12 年。配偶与受访者的受教育水平基本保持一致。married 数值为 0.987，样本基本上都属于已婚状态，波动非常小。该样本表明本次研究的问题是都是在已婚受访者的基础上进行的问题研究。其中学龄前儿童，和 7-18 儿童的均值基本为 1，且波动幅度很小，另外健康以及 female 的均值和波动同样小，而 exper 的均值为 11，波动值非常大。这些因素对工资的影响是否可以说明对线性回归显著性的影响，这将是我们通过分析进行讨论说明的问题。

模型建立

模型一：教育程度对工资影响

$$wage = \alpha_1 + \beta_1 edu + \beta_2 female + \gamma X + \varepsilon$$

模型二：教育对工资对数影响

$$lwage = \alpha_2 + \beta_3 edu + \beta_4 female + \gamma_1 X + \varepsilon$$

模型三：性别在教育对工资影响中的作用

$$lwage = \alpha_3 + \beta_5 edu + \beta_6 female + \beta_7 femedu + \gamma_2 X + \varepsilon$$

数据分析

	(1)	(2)	(3)

VARIABLES	wage	lwage	lwage
age	0.909	-0.00223	-0.00187
	(0.558)	(0.00224)	(0.00224)
edu	6.474***	0.0722***	0.0838***
	(1.291)	(0.00693)	(0.00779)
femedu			-0.0260***
			(0.00800)
kidsum	8.771	0.0572	0.0530
	(13.43)	(0.193)	(0.199)
childrensum	-9.518	-0.114	-0.112
	(13.60)	(0.191)	(0.197)
female	20.06***	0.215***	0.554***
	(6.429)	(0.0275)	(0.107)
health	6.618	0.135***	0.133***
	(7.472)	(0.0288)	(0.0288)
exper	-3.942*	-0.00566	-0.00583
	(2.088)	(0.00547)	(0.00548)
exper2	0.106*	0.000355**	0.000345*
	(0.0591)	(0.000176)	(0.000177)
spousewage	0.00122**	9.28e-06**	9.58e-06**
	(0.000616)	(3.85e-06)	(3.87e-06)
spouseedu	3.440***	0.0251***	0.0264***
	(1.109)	(0.00640)	(0.00639)
Constant	-67.12**	2.765***	2.593***
	(29.65)	(0.120)	(0.129)
Observations	2,902	2,895	2,895
R-squared	0.031	0.223	0.226

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

模型一：教育程度对工资影响

$$wage = 6.474edu + 20.06female + \gamma X - 67.12$$

$$wage = 0.909age + 6.474edu + 8.771kidsum - 9.518childrensum$$
$$+20.06female + 6.618health - 3.942\exp er + 0.106\exp er2$$
$$+0.00122spousewage + 3.440spouseedu - 67.12$$

模型二：教育对工资对数影响

$$lwage = 0.0722edu + 0.215female + \gamma_1 X + 2.765$$

$$lwage = -0.00223age + 0.0722edu + 0.0572kidsum$$
$$-0.114childrensum + 0.215female + 0.135health$$
$$-0.00566\exp er + 0.000355\exp er2 + 9.28e\hat{}(-06)spousewage$$
$$+0.0251spouseedu + 2.765$$

模型三：性别在教育对工资影响中的作用

$$lwage = 0.0838edu + 0.554female - 0.0259femedu + \gamma X + 2.59$$

$$lwage = -0.00187age + 0.0838edu + 0.0529kidsum$$
$$-0.1124hildrensum + 0.554female + 0.133health - 0.0259famedu$$
$$-0.00583\exp er + 0.000345\exp er2 + 9.58e\hat{}(-06)spousewage$$
$$+0.0264spouseedu + 2.59$$

注：这里，我们只对不为缺失值的变量做了回归

模型 1，小时工资方程

female 的系数度量的是在给定相同水平 educ 和其他变量时，一个女性和一个男性在小时工资上的平均差距。如果我们找到受教育程度、工作经历等相同的一个男性和女性，那么平均来看，女性每小时比男性少挣 20.06。由于我们已经进行了多元回归，并控制了 X，所以这 20.06 的工资差距不能由男女在 X 上的平均差距来解释。我们可以断定，这 20.06 的差别是有性别或者我们在回归中没有控制的与性别相关的因素所导致的。

模型 2，对数小时工资方程

我们将模型 1 的工资方程的因变量换成 log（wage），并且通过增加 X 的二次项重新估计它。female 的系数意味着，在 X 相同水平上，女性比男性约少挣 100*0.215=21.5%。模型 2 在原有模型 1 基础上，改变因变量为 log（wage），原因在于。首先，当因变量 y 大于 0,时，使用 log（y）作为因变量的模型，通常比使用 y 的水平值作为因变量的模型更接近 CLM 假定。严格为证的变量，其条件分布常常具有异方差性或偏态性，取对数后，即使不能消除两方面的问

题，也可以使之有所缓和。我们可以在 figure 中看到，模型 2 中的结果显著性要由于模型 1 的显著性。

模型 3，对数小时工资方程（存在交互项）

我们关心的是教育的回报，从报告的系数来看，可以看出教育每上升一年，每小时工资对数将上升 0.0838 元。从 p 值来看，edu、female、health、spousewage 是显著异于 0 的，及从统计上将对 wage 有解释作用。该模型在模型 2 的基础上加了 femedu 这一交互项，通过 figure 可以看出，当教育水平取均值 12.95 时，女性的收入比男性的收入会减少 0.0259。

在三个模型中，教育，性别为男性，配偶的教育程度对受访者的工资都有 1%的显著为正的影响。配偶的工资对受访者工资有 5%显著为正的影响,但影响很小。而家庭拥有学龄前和学龄期孩子的个数则对受访者工资没有显著的影响。

此外，其中学龄前儿童，和 7-18 儿童的均值基本为 1，且波动幅度很小，他们对工资的影响是不显著的。另外健康以及 female 的均值和波动同样小，二者对工资影响显著。而 exper 的均值为 11，波动值非常大，但却并不显著。我们可以看出均值和波动值大小不是直接体现显著性差异的特征。