2008 年 5 月 数理统计与管理 May 2008
第 27 卷 第 3 期 Application of Statistics and Management Vol. 27 No. 3

文章编号：1002-1566(2008)03-0404-05

教育收益率估计方法的比较

王明进 陈良焜

(北京大学光华管理学院，北京 100871)

摘要：本文对教育经济学中个人教育投资明瑟收益率的估计方法进行了比较研究，特别是引入了部分线性回归模型对明瑟收益率进行估计的方法，并且通过广义似然比检验对常规明瑟模型中的参数形式进行了检验。

关键词：教育投资；明瑟收益率；部分线性回归；广义似然比

中图分类号：G40-054; O212 **文献标识码**：A

Comparison of Different Estimation Methods for Rates of Return to Education

WANG Ming-jin, CHEN Liang-kun

(Guanghua School of Management, Peking University, Beijing 100871, China)

Abstract: This paper compares the different estimation methods for the Mincerian private rate of return to education. In particular, we put forward a new estimation method based on Partially Linear Regression (PLR) model and compare it with the traditional Mincer's regression model with the quadratic form for the working year variable. Generalized likelihood ratio test is conducted to verify the parametric form of the Mincer's model which is widely used in the literatures.

Key words: education investment, Mincerian rate of return, partially linear regression, generalized likelihood ratio

0 引言

对个人教育投资收益的研究是教育经济学里面的一个重要内容。这种收益是指受教育者自身及其家庭从接受教育中所获得的收益，它既包括了货币收益，也包括了非货币收益，前者体现在受教育者货币收入的提高，后者则体现在受教育者素质的提高等多个方面。在教育经济学中，人们更加关注的是个人教育的货币收益。将接受教育视为一个投资过程，为了度量其收益率既可以使用内部收益率，也可以使用所谓的明瑟收益率 [1]。近些年来，国内的很多学者利用调查数据对近 20 年来我国城镇教育的收益率及其变化进行了大量的实证研究，具体可以参见 [2-5] 等。这些研究大多数考虑的都是教育的明瑟收益率。

如果用 y 来表示就业者的收入，用 x_1 来表示其接受教育的时间（以年计算），那么明瑟收益率可以定义为 [1]：

$$\gamma = \Delta \ln y / \Delta x_1, \tag{1}$$

它度量了一个就业者每多接受一年的教育，其收入增加的比例是多少。通常明瑟收益率的估计可以通过建立如下的明瑟回归模型而得到，

$$\ln y = \beta_0 + \beta_1 x_1 + \delta_1 z + \delta_2 z^2 + \varepsilon. \tag{2}$$

这里的 z 表示就业者的工龄，而系数 β_1 就是前面定义的明瑟收益率 γ。普通的最小二乘估计给出了明瑟收益率的估计值。按照我们在 [5] 中的讨论，这里的明瑟收益率其实是一种平均意

收稿日期：2007 年 10 月 26 日

义上的收益率，即不同就业者多接受一年教育其收入增加的比例的平均值。当然，在估计明瑟收益率时，也可以将就业者的除了工龄之外的其他特征作为控制变量考虑进来。比如，在本文中，我们还考虑了就业者的性别 x_2、所处的地域 x_3 等影响其收入的因素，因此采用的是明瑟回归模型如下的一种扩展形式：

$$\ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \delta_1 z + \delta_2 z^2 + \varepsilon. \tag{3}$$

明瑟收益率 γ 的估计仍然由 β_1 的最小二乘估计给出来。这与前面给出的几个文献中的做法是一致的。

应该说，尽管如此常用，但是在明瑟回归模型中将 "工龄" 项设置成二次曲线的形式只是一种简化或者一种经验的反映，即工龄对收入的影响不应该是简单的直线形式。从理论上讲，如果工龄的影响的确是曲线形式，即 $\delta_2 \neq 0$，那么仅仅考虑 z 的一次形式得到的 β_1 的最小二乘估计将不再是明瑟收益率的一个无偏的估计，当然，估计的精度（如果以均方误差来衡量的话）未见得就一定会降低。具体这方面的讨论可以参照文献 [6]。

在实证研究中是否 "工龄" 项的二次曲线形式就是一种合适的设定，或者这种设定对明瑟收益率的估计会造成什么样的结果，这是本文将要关注和研究的问题。为此，本文提出利用一个半参数模型，即部分线性回归模型，来估计教育的明瑟收益率。在这个模型中对 "工龄" 项的具体形式没有进行设定，而是需要根据数据一起估计出来。特别地，我们还通过一种广义似然比检验来对二次曲线的设置形式进行了判断，并且分析不同设置形式会对明瑟收益率的估计造成什么样的影响。这种比较和分析对于教育经济学研究的意义不言而喻，而在同类文献中使用非参数、半参数的模型则完全是一种新的尝试。

下面各个部分的主要内容包括：第 1 节介绍了估计明瑟收益率的半参数模型；第 2 节是所用数据的一个说明，第 3 节给出了半参数模型和常规明瑟模型的比较结果；第 4 节是结论。

1 部分线性模型

为了估计明瑟收益率，本文采用的部分线性模型（Partial Linear Model，简记作 PLM）的形式为

$$\ln y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + f(z) + \varepsilon, \tag{4}$$

其中自变量 x_1, x_2, x_3, z 的含义同前，与明瑟回归模型 (3) 相比，这里工龄对收入的影响 $f(z)$ 形式是未知的，通常只假定它是一个光滑函数，因此是一种非参数的形式，而 x_1, x_2, x_3 对收入的影响仍然是线性的。上述模型也被称为部分线性模型，属于半参数模型的一种。关于部分线性模型的文献可以参见 [7–10] 等。容易理解，模型 (4) 中的系数 β_1 也是教育的明瑟收益率。

给定样本 $\mathbf{x}_i = (x_{1i}, x_{2i}, x_{3i})'$，$z_i$，$y_i$，$i = 1, 2, \cdots, n$，为方便，可以将模型 (4) 写成如下的矩阵形式，

$$\mathbf{y} = \mathbf{X}\beta + f(\mathbf{z}) + \varepsilon, \tag{5}$$

其中 $\mathbf{y} = (\ln y_1, \ln y_2, \cdots, \ln y_n)'$，$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$，$f(\mathbf{z}) = (f(z_1), f(z_2), \cdots, f(z_n))'$，$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$，$\beta = (\beta_1, \beta_2, \beta_3)'$.

估计系数 β 和 $f(\mathbf{z})$ 的方法有多种。比如，惩罚最小二乘估计方法（the penalized least squares method; [7–8]）、部分回归方法（the partial regression approach; [9–10]）等。这里我们采用的是如下形式的部分回归估计：

$$\begin{aligned} \widehat{\beta} &= (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{y}}, \\ \widehat{f}(\mathbf{z}) &= \mathbf{W}(\mathbf{y} - \mathbf{X}\widehat{\beta}), \end{aligned} \tag{6}$$

其中的 $\widetilde{\mathbf{y}} = (\mathbf{I}\text{-}\mathbf{W})\mathbf{y}$，$\widetilde{\mathbf{X}} = (\mathbf{I}\text{-}\mathbf{W})\mathbf{X}$，而 $\mathbf{W}$ 是所用的平滑方法对应的 $n \times n$ 的转换矩阵，又称为平滑矩阵，类似于线性回归模型中的帽子矩阵。Speckman[9] 给出了 $\mathbf{W}$ 是核函数平滑矩阵时估计 $\widehat{\beta}$ 和 $\widehat{f}(\mathbf{z})$ 的渐近性质；Hamilton[10] 则给出了是局部线性平滑矩阵时 $\widehat{\beta}$ 和 $\widehat{f}(\mathbf{z})$ 的渐近分布。考虑到局部线性估计比核估计和样条估计都有更好的性质，因此本文采用的都是局部线性估

计 [11]，平滑矩阵为 $\mathbf{W}=(w(z_i,z_j))_{n\times n}$，

$$w_h(z_i,z)=\frac{\big(s_{2,h}(z)-s_{1,h}(z)(z_i-z)\big)k_h(z_i-z)}{s_{0,h}(z)s_{2,h}(z)-s_{1,h}(z)^2},$$

其中的 $s_{l,h}(z)=\sum_{j=1}^{n}(z_j-z)^l k_h(z_j-z)$, $l=0,1,2$. $k(u)$ 是核函数，$k_h(u)=(1/h)k(u/h)$，h 是带宽。此时估计的渐近分布则可以参照 [10]。

为了判断明瑟回归模型中 (3) 中对工龄项设置的二次曲线形式是否合理，可以通过检验假设 H_0：明瑟回归模型 (3) 成立。对应的备择假设是将工龄项设置为二次曲线形式不合适，而应该是其他形式的曲线，具体来说，H_1：部分线性模型 (4) 成立，当然其中的 $f(z)$ 不是二次曲线。

为了检验 H_0，这里采用广义似然比检验 (参见 [12])，检验统计量为

$$T=\frac{n}{2}\log(RSS_0/RSS_1), \tag{7}$$

其中 RSS_0 是对数据拟合明瑟回归模型后得到的误差平方和，RSS_1 是由部分线性模型拟合得到的误差平方和。类似于 [12]，该检验的 p- 值可以通过如下的条件 Bootstrap 方法得到。具体步骤为：

1) 用 $\{\widehat{\varepsilon}_i\}$ 记对数据拟合部分线性模型后得到的残差。在 $\{\widehat{\varepsilon}_i-\bar{\widehat{\varepsilon}}_i\}_1^n$ 中可重复地抽取 ε_i^*, $i=1,2,\cdots,n$，构造条件 Bootstrap 样本

$$\ln y_i^*=\widehat{\beta}_0+\widehat{\beta}_1x_{i1}+\widehat{\beta}_2x_{i2}+\widehat{\beta}_3x_{i3}+\widehat{\delta}_1z_i+\widehat{\delta}_2z_i^2+\varepsilon_i^*,$$

其中 $i=1,2,\cdots,n$。这里的系数为对数据拟合明瑟回归模型得到的系数估计值。

2) 利用 Bootstrap 样本 $\{\mathbf{y}^*,\mathbf{X},\mathbf{z}\}$ 构造广义似然比统计量 T^*；

3) 重复上述 1)、2) 步骤 B 次，得到 B 个统计量的值 $T_1^*,T_2^*,\cdots,T_B^*$；

4) 检验的 p- 值则是 $T_1^*,T_2^*,\cdots,T_B^*$ 中大于 T 的比例。

2 数据说明

本文使用的数据来自国家统计局 1991 年、1995 年、2000 年以及 2004 年的中国城市住户调查数据，所考察的变量包括被调查者的工资收入 (y)、接受教育的年数 (x_1)、性别 (x_2)、所在地区 (x_3)、工龄 (z) 等。由于原始数据中包含了个别明显的错误以及变量的含义也并不完全一样，因此，我们事先对数据做了一定的处理，比如去除了数据中明显的不合理的观测，包括性别、学历、就业、省份等不在正确的编码范围内的情况；去除了工龄不到 1 年 (即 0 年) 以及年龄不满 16 岁或超过 70 岁的观测；去除了全部数据中工资收入为零的观测等。另外，被调查者接受教育的年数则按照其最高学历进行了转换，具体包括：研究生学历记为 18.5 年，大学本科学历为 15.5 年，大学专科为 14.5 年，中专和高中学历均记为 11.5 年，初中学历为 8.5 年，小学学历为 5.5 年，其它未上学或者只进过扫盲班的为 0 年。另外，由于目前得到的 1995 年的数据中没有性别因素，而且数据中剔除了收入低于 1000 元的被调查者，因此对 95 年结果的解读需要考虑到这些影响。其他对数据处理的细节可以参考 [5]。

表 1 对样本数据的简单描述

	91 年	95 年	00 年	04 年
样本大小 n	5000	5000	5000	5000
收入平均值	2661	6752	9418	15619
收入中位数	2477	5956	8124	12546
收入标准差	1349	4123	6531	13076
女性人数	2394		2280	2277

考虑到估计半参数模型 (4) 的算法比较费时，我们对前面处理过的原始数据进行了随机抽样，每年数据的样本都由 5000 人组成。对四个样本的统计描述结果由图 1 给出。

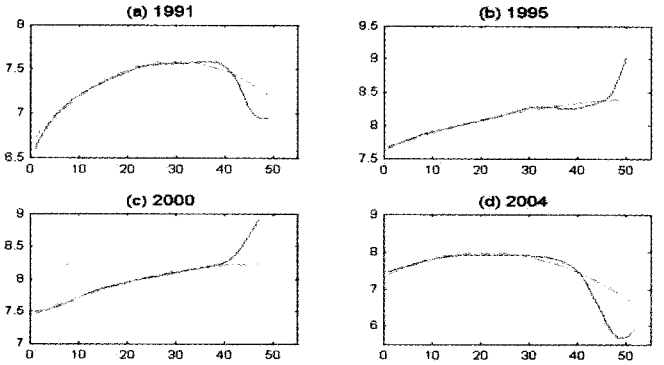

图 1 工龄对收入的影响曲线，其中实线是半参数模型估计的结果（平滑系数均按照 0.3 倍工龄的样本标准差），虚线是二次曲线估计的结果。

3 模型估计的结果分析

为了比较对工龄项的不同设置形式对估计明瑟收益率所造成的影响，我们分别考察了不包括工龄的项（即在 (3) 中去掉 z 和 z^2 项）、工龄项的线性形式（即在 (3) 中去掉 z^2 项）、二次形式（即模型 (3)）以及非参数形式（即模型 (4)）所对应的不同模型。在利用局部线性方法对模型 (4) 进行估计时，尽管平滑系数的选择原则上可以通过最小化 GCV 函数 [9] 获得，但是依照我们的经验，直接利用最小化 GCV 函数得到的窗宽往往都太小，最终估计的曲线都过于曲折，因此并不适合识别其形态。这里通过尝试比较几种不同选择，将平滑系数选为 0.3 倍的工龄变量的样本标准差。

表 2 明瑟收益率估计结果的比较

	模型	收益率估计	标准误差	模型的 R^2	模型的 $\widehat{\sigma}^2$
91 年	不考虑	0.0203	0.0025	0.0699	0.2561
	线性	0.0297	0.0023	0.2313	0.2117
	二次	0.0296	0.0022	0.2780	0.1989
	半参数	0.0294	0.0022	0.2837	0.1971
95 年	不考虑	0.0374	0.0027	0.1188	0.2463
	线性	0.0444	0.0026	0.2003	0.2235
	二次	0.0446	0.0026	0.2021	0.2231
	半参数	0.0445	0.0026	0.2046	0.2222
00 年	不考虑	0.0739	0.0038	0.1309	0.4763
	线性	0.0869	0.0037	0.2005	0.4383
	二次	0.0864	0.0037	0.2027	0.4372
	半参数	0.0862	0.0037	0.2041	0.4360
04 年	不考虑	0.1225	0.0044	0.1913	0.7176
	线性	0.1236	0.0046	0.1915	0.7176
	二次	0.1199	0.0045	0.2263	0.6868
	半参数	0.1197	0.0044	0.2421	0.6720

针对四年的数据利用上述三种形式得到的明瑟收益率及估计误差在表 2 中给出。我们注意到，相比工龄为线性形式的回归模型，将工龄的影响用二次曲线表达的模型得到的明瑟收益率的估计要稍微小一些（只有 95 年结果是个例外）；而且，相比工龄影响为二次曲线的回归模型，部分线性模型给出的明瑟收益率的估计又都略微小了一点。当然，这里得到的明瑟收益率的估计值也与文献 [2–4] 中使用二次曲线形式所得到的对应结果是很接近的。

根据表 2 给出的模型的 R 平方和误差项方差的估计结果，半参数模型均比线性和二次的模型对数据拟合得更好。图 1 中的实线给出了模型 (4) 中估计得到的工龄对收入的影响曲线，即 $\widehat{f}(z)$，而虚线则给出了利用模型 (3) 估计出的影响曲线，即 $\widehat{\beta}_0+\widehat{\delta}_1 z+\widehat{\delta}_2 z^2$. 不难看出，四年数据给出都是凹的二次曲线，即 $\widehat{\delta}_2<0$，而且，对比两种模型的结果，对于工龄小于 30 年的情况，二次曲线基本上都比较好地描述了工龄对收入的影响；但是，当工龄超过 30 多年的时候，工

龄对收入的影响就变得比较复杂了。在 91 年和 04 年的样本数据中，这部分人的收入都偏低，而 95 年和 00 年的情形似乎正好相反，这部分人的收入反而是很高的。在四组样本中，工龄超过 30 年的人数所占的比例分别为 1991 年 11.7%，1995 年 11.3%，2000 年 13.6%，2004 年 20.2%。

表 3 广义似然比检验的结果

年份	91	95	00	04
统计量的值	19.883	7.763	4.392	51.584
p-值 *	0.002	0.040	0.217	0.000

* 注：这个的 $p-$ 值由条件 Bootstrap 得到，$B = 1000$.

表 3 给出了对四组样本检验 "工龄因素的影响符合二次曲线形式" 的假设所得到的广义似然比统计量以及由条件 Bootstrap 方法得到的 $p-$ 值。与图 1 给出的印象一致，对于 91、95、04 年的样本数据都有比较充分的证据显示 "工龄的二次曲线" 形式并不是一个合适的表达方式。只有对 2000 年的一组样本才显示这种证据不足。

当然，从明瑟收益率及估计误差的取值大小上看，考虑工龄因素后的三种模型得到的结果其实都是比较接近的。这显示了对 "工龄" 项的具体形式的设置对于估计明瑟收益率并不是一个很大的问题。相反，不考虑工龄因素得到的明瑟收益率则与前面分析的结果呈现出一定的差距。特别是在前三年的结果中，不考虑工龄因素明显低估了教育的明瑟收益率。

4 结论

本文对估计教育明瑟收益率的不同方法进行了比较研究，特别是就 "工龄" 项的不同的设定形式对明瑟收益率估计所造成的影响进行了分析。我们特别引入了估计明瑟收益率的部分线性回归模型，发现通常文献中使用的二次曲线并不能正确表达工龄对收入的影响，特别是对工龄较长的情形，二次曲线和非参数形式得到的结果差距会很大。值得庆幸的是，工龄影响的形式的不同设置对明瑟收益率的估计值并没有造成多大的影响，换句话说，即便不考虑工龄的二次项也可以得到非常接近的结果。但是，去掉工龄因素，往往会低估明瑟收益率。显然，上述结论对教育收益率的实证研究有一定的参考价值。

[参考文献]

[1] Mincer, J. Schooling, Experience and Earnings [M]. New York: National Bureau of Economic Research, 1974.

[2] 陈晓宇、陈良焜、夏晨. 20 世纪 90 年代中国城镇教育收益率的变化与启示 [J]. 北京大学教育评论, 2003, (4).

[3] 马晓强、丁小浩. 我国城镇居民个人教育投资风险的实证研究 [J]. 教育研究, 2005, (4): 25–31.

[4] 岳昌君、刘燕萍. 教育对不同群体收入的影响 [J]. 北京大学教育评论, 2006, (2): 85–92.

[5] 王明进、岳昌君. 个人教育投资风险的计量分析 [J]. 北京大学教育评论, 2007 (2): 128–135.

[6] 陈希孺、王松桂. 近代回归分析 [M]. 合肥：安徽教育出版社, 1987.

[7] Engle, R., Granger, C.,Rice, J. Weiss, A. Nonparametric estimates of the relation between weather and electricity sales [J]. Journal of American Statistical Association , 1986, (81): 310–320.

[8] Green, P. Linear models for field trials, smoothing and cross-validation [J]. Biometrika , 1986, (72): 527–537.

[9] Speckman, P. Kernel smoothing in partial linear models [J]. Journal of Royal Statistical Society B, 1988, (3): 413–436.

[10] Hamilton, S. A. Local linear estimation in partly linear models [J]. Journal of Multivariate Analysis, 1997, (60): 1–19.

[11] Fan, J., Gijbels, I. Local Polynomial Modeling and Its Applications [M]. London: Chapman & Hall, 1996.

[12] Fan, J., Zhang, C., Zhang, J. Generalized likelihood test statistic and the Wilks phenomenon [J]. The Annals of Statistics, 29: 153–193.