Econometrics

01 单变量回归模型

1.1 引言

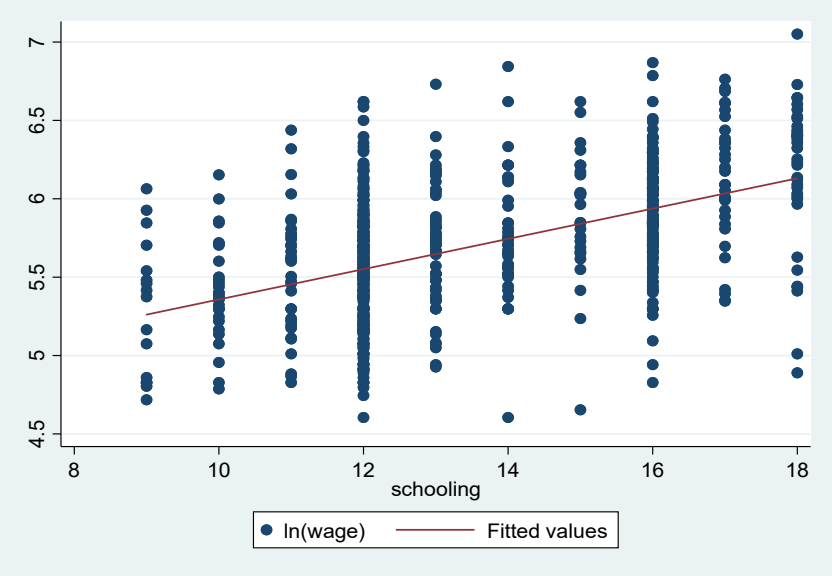
“回归（regression）”的由来

由生物学家高尔顿引入，他发现虽然身高存在一定的趋势，即父母高子女也高，但是在给定父母身高的同时，子女的平均身高却趋向于或“回归”到全体人口的平均身高。即对于一个父辈高的群体，子女的平均身高低于父辈身高，而对于一个父辈矮的群体，子女的身高则高于父辈的身高，整体而言“回归到中等”。

回归的现代含义

研究一个所谓的因变量对另一个或者多个所谓自变量的依赖关系，目的是通过后者（在重复抽样过程中）的已知或设定值，去估计或预测前者的（总体）均值。

【例】考察工资对数与教育年限之间的线性关系，利用 grilic.dta



统计关系与确定性关系

由上文的例子可以看出，由于误差项的存在，变量之间的关系是一种统计依赖关系。不同于自然科学中的确定性依赖关系，我们主要处理的是**随机**（stochastic 或 random）变量，即有概率分布的变量。

统计依赖关系存在的原因在于变量的测量可能有误差，还有很多影响因变量的因素无法一一辨认出来，在工资决定方程之中，能力，家庭背景等很多因素都有可能影响最终的工资决定。随机变量的存在意味着我们需要靠数理统计的相关知识来进行分析。

回归的注意点：

- 回归分析研究一个变量对于另一个（或一类）变量的依赖关系，但并不一定意味着因果关系，即统计关系并不意味着任何因果关系。（因果识别是现代计量经济学最大的特点）
- 回归分析并不等同于相关分析，回归分析的因变量与自变量并不对称，回归分析的因变量是随机的，自变量是固定的。而相关分析对称地对待任何变量。

经济分析所用数据的类型

- **横截面数据**（cross-section data）：同一时间点上收集的数据。
- **时间序列数据**（time series data）：一个变量在不同时间取值的一组观测结果。
- **混合数据**（pooled data）：横截面数据+时间序列数据。
- **面板数据**（panel data）：对相同的横截面单位在时间轴上进行跟踪调查的数据。

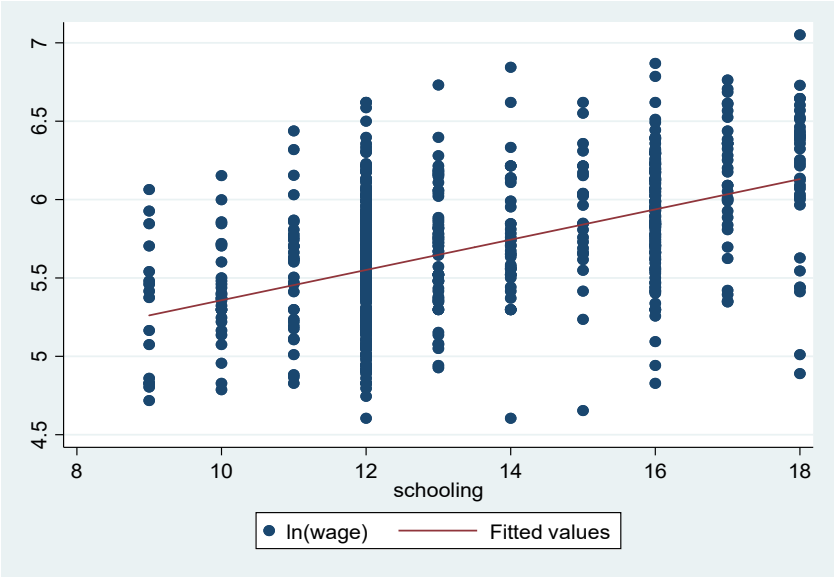
1.2 一个回归分析的例子

回到上文工资的对数与工作年限的关系的例子，教育年限有 9-18 年，每个确定的教育年限有不同的工资分布，但是随着教育年限的增加，工资对数的平均值也在增加。在每个教育年限都对应一个均值，这个均值成为**条件期望值**（conditional expected values），或称为条件均值，记作

$E(Y|X)$，即为给定 X 条件下 Y 的期望值。而**无条件期望值**（unconditional expected values）为整个总体的期望值，记作 $E(Y)$。

回归分析的目的是进行预测，当我们想得到教育年限在 12 年的人的工资期望值是多少时，应该回答的是当 s=12 时候的条件均值。

将所有的条件均值连接起来，就得到所谓的**总体回归线**（population regression line，PRL），就是 ***Y* 对 *X* 的回归**。



则在几何意义上，PRL 就是当解释变量取给定值的时的因变量的条件均值的轨迹。

总体回归函数的概念

综上，我们的总体回归线是给定给定 X 下的条件期望的均值的轨迹，而每一条件期望都是 X 的一个函数，用符号表示为

$$E(Y|X_i) = f(X_i)$$

即总体回归函数，它给出了 Y 的均值是如何随着 X 的变化而变化的。总体回归函数的设定是一个经验方面的问题，需要经济理论的指导。

假定为线性函数：

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

其中β_1被称为截距（intercept），β_2被称作斜率（slope），该方程被称为总体回归函数（PRF）。

线性的含义

计量经济学中的线性回归值得是相对于参数是线性的，即 $\mathrm{E}(Y|X_i)=\beta_1+\beta_2X_i^2$是线性的，而 $\mathrm{E}(Y|X_i)=\beta_1+\frac{\beta_2}{X_i}$就不是线性的。

PRF 的随机设定

由这个例子的图可以看出，在给定教育年限的条件下，工资对数分布在其均值周围，因此可以将工资决定方程表述为：

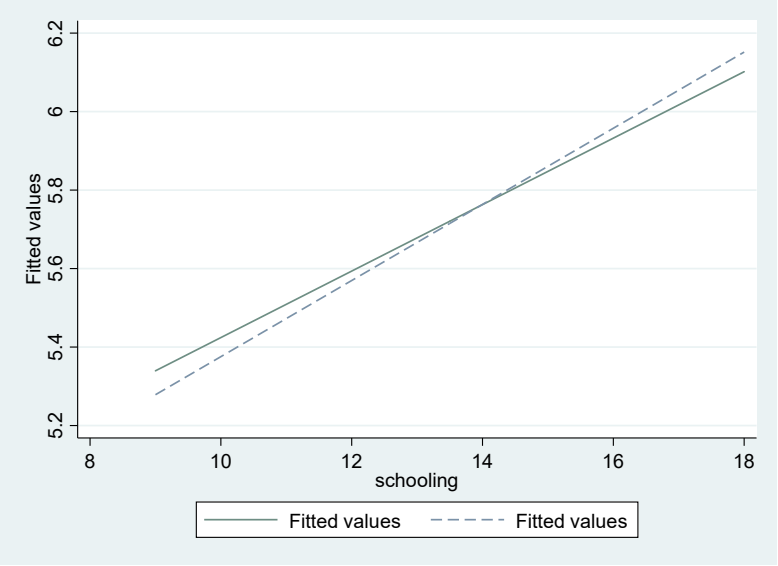
$$Y_i=E(Y|X_i)+u_i$$

则可以将给定的教育年限下的工资对数分成两部分，一部分系统性成分，另一部分非系统性成分，由随机误差项u_i(stochastic disturbance term)组成。将上式两边同时取期望可得到：$E(u_i|X_i)=0$。因此假定回归线通过 Y 的条件均值意味着u_i的条件均值就是零。

样本回归函数

现实情况中，我们面临的是只有总体的某一个样本，因此目的是在样本信息的基础上来估计 PRF。

假装不知道工资与教育年限的数据，我们从中随机抽取两个各 100 个个体的样本。由于抽样波动，我们未必能准确计算出 PRF，将两个样本的散点进行拟合得到两条**样本回归线**（sample regression line,SRL），可以看出差距并不是很大，姑且认为是真实 PRL 的一个近似。



可以计算出样本回归函数（SRF）

$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} X_i$$

其中 $\widehat{\beta_1}$、$\widehat{\beta_2}$被称为估计量（estimator），计算出来的为估计值，表达为随机形式为：

$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} x_i + u_i$$

则回归分析的目的就是通过 SRF 来估计 PRF。需要使得估计量尽可能的接近真实值β_1、β_2，尽管这个真实值可能永远都不知道。