

Stata 基础训练营

第三讲 回归分析与结果输出

江俊澄

jc_jiang@stu.zuel.edu.cn

中南财经政法大学财政税务学院
学生学术研究中心 副主编

2024.12.01



回顾：数据分类

通常来说，可以将经济数据分为三/四类：

- ▶ **横截面数据 cross-section data**
 - ▶ 由在给定时间点采集的个人、家庭、公司、城市、国家或各种其他单位的样本组成。
 - ▶ 忽略微小的时差，在一年内不同月份对不同的人进行了调查。
- ▶ **时间序列数据 time series data**
 - ▶ 由对一个或多个变量随时间变化的观察组成。
 - ▶ 很少假设经济数据在时间上是独立的。
 - ▶ 需要关注数据搜集时的数据频率。(每天/周/月/季/年？)
- ▶ **(混合横截面数据 pooled cross section)**
 - ▶ 同时具有横截面和时间序列特征。
如：2022 年调查了一组人，2024 年调查了新的一组人。
- ▶ **面板数据 panel data**
 - ▶ 由数据集中每个横截面单位的一个时间序列组成。

回归: regression

回归的古典意义:

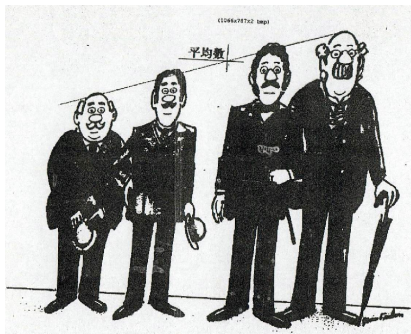
高尔顿遗传学的回归概念 (父母身高与子女身高的关系) 子女的身高有向人的平均身高“回归”的趋势

回归的现代意义:

一个被解释变量对若干个解释变量依存关系的研究

回归的目的 (实质):

由解释变量估计被解释变量的平均值



OLS: Ordinary Least Squares

以一元线性回归为例，有如下：

$$y = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, 2, 3 \dots n)$$

- ▶ y_i : 被解释变量 (dependent variable, regressand)
- ▶ x_i : 解释变量 (explanatory variable, independent variable, regressor)
- ▶ α : 截距项 (intercept) 或常数项 (constant);
- ▶ β : 斜率 (slope)
- ▶ α 与 β : 统称“回归系数” (regression coefficients) 或“参数” (parameters)。
- ▶ ε_i : “误差项” (error term) 或“扰动项” (disturbance), 包括遗漏的其他因素、变量的测量误差、回归函数的设定误差 (比如, 忽略了非线性项) 以及人类行为的内在随机性等。

OLS: 普通最小二乘法原理

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

残差平方和最小:

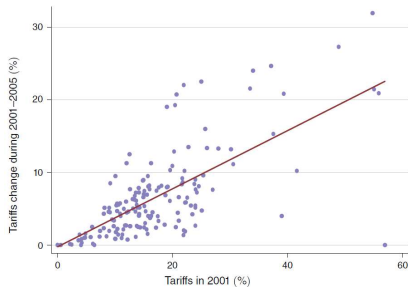
$$\min \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i})^2$$

样本回归函数为:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

i.e.

$$y_i = \hat{y}_i + \hat{\varepsilon}_i$$



R^2 的得来

分析 Y 的观测值 Y_i 、估计值 $\hat{Y}_i$ 与平均值 $\bar{Y}$ 有以下关系

$$Y_i - \bar{Y} = (Y_i - \bar{Y}) + \hat{Y}_i - \hat{Y}_i = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

将上式两边平方加总，可证得 (提示：交叉项 $\sum (\hat{Y}_i - \bar{Y})e_i = 0$)

$$\sum_{(TSS)} (Y_i - \bar{Y})^2 = \sum_{(ESS)} (\hat{Y}_i - \bar{Y})^2 + \sum_{(RSS)} (Y_i - \hat{Y}_i)^2$$

或者表示为 $\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$

- ▶ 总变差 $\sum y_i^2$ (TSS): 被解释变量 Y 的观测值与其平均值的离差平方和 (总平方和) (说明 Y 的总变动程度)
- ▶ 解释了的变差 $\sum \hat{y}_i^2$ (ESS): 被解释变量 Y 的估计值与其平均值的离差平方和 (回归平方和)
- ▶ 剩余平方和 $\sum e_i^2$ (RSS): 被解释变量观测值与估计值之差的平方和 (未解释的平方和)

何为 R^2 !

什么是 R 方?

定义: 回归平方和 (解释了的变差 ESS) $\sum \hat{y}_i^2$ 在总变差 (TSS) $\sum y_i^2$ 中所占的比重称为可决系数, 用 r^2 或 R^2 表示:

$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} \quad \text{或} \quad R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

取值在 0-1, 是是非负的统计量, 也被称作: 拟合优度。说明的是模型对样本观测值的拟合程度。

不过, R^2 并不是特别重要, 一般也不在文章中报告 qwq
更重要的是 t 值和 p 值。

回归系数标准误

在回归模型中，回归系数估计值的标准差一般称为标准误 (standard error)。

回归系数标准误，用于衡量回归系数估计值的精确性和稳定性。标准误越小，说明回归系数的整体精度越高。或言之，是样本回归系数分布的标准差。

有一个关系式，大家在报告回归分析结果结论时需要知道。

$$\frac{\text{回归系数}}{\text{标准误}} = t\text{值}$$

根据 t 值或 t 分布就可以获得某个自变量因素对 Y 影响的显著性 P 值。

所以，我们在论文中常见的格式是：回归系数 (se)。
利用回归系数的标准误，可以构造出该系数的置信区间。置信区间提供了一个范围，该范围内的值在统计上被认为是可能的真实值。

回归系数的检验方法:z 检验

确立假设: 原假设为 $H_0: \beta_2 = 0$

备择假设为 $H_1: \beta_2 \neq 0$

(本质: 检验 β_2 是否为0, 即检验 X_i 是否对 Y 有显著影响)

(1) 当已知 σ^2 或样本容量足够大时

可利用正态分布作Z检验

$$Z^* = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim N(0,1)$$

给定 α , 查正态分布表得临界值 Z

▼ 如果 $-Z < Z^* < Z$ 则不拒绝原假设 H_0

▼ 如果 $Z^* < -Z$ 或 $Z^* > Z$ 则拒绝原假设 H_0

回归系数的检验方法:t 检验

(2) 当 σ^2 未知, 且样本容量较小时

只能用 $\hat{\sigma}^2$ 去代替 σ^2 , 可利用 t 分布作 t 检验:

$$t^* = \frac{\hat{\beta}_2 - \beta_2}{\hat{SE}(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{\hat{SE}(\hat{\beta}_2)} \sim t(n-2)$$

给定 α , 查 t 分布表得 $t_{\alpha/2}(n-2)$

▼ 如果 $t^* \leq -t_{\alpha/2}(n-2)$ 或者 $t^* \geq t_{\alpha/2}(n-2)$

则拒绝原假设 $H_0: \beta_2 = 0$ 而不拒绝备择假设 $H_1: \beta_2 \neq 0$

▼ 如果 $-t_{\alpha/2}(n-2) \leq t^* \leq t_{\alpha/2}(n-2)$

则不拒绝原假设 $H_0: \beta_2 = 0$

用 **P** 值判断参数显著性的方法

方法： 将给定的显著性水平 α 与 **p** 值比较：

- ▶ 若 $\alpha > p$ 值，必有 $|t^*| > |t_{\alpha/2}|$ ，则在显著性水平 α 下拒绝原假设 $H_0: \beta_k = 0$ ，即认为 **X** 对 **Y** 有显著影响
- ▶ 若 $\alpha \leq p$ 值，必有 $|t^*| \leq |t_{\alpha/2}|$ ，则在显著性水平 α 下不拒绝原假设 $H_0: \beta_k = 0$ ，即认为 **X** 对 **Y** 没有显著影响

规则： 当 $p < \alpha$ 时，**P**值越小，越能拒绝原假设 H_0

几种形式的系数含义

对于任何基本的线性回归，在数据处理有时拿原值直接进行回归，有时会做对数化处理，如下是常见（对数 × 线性）的组合方式，不同形式下对 β_1 的解读不同。

$$y = \beta_1 x + \beta_0 + \varepsilon$$

$$y = \beta_1 \ln x + \beta_0 + \varepsilon$$

$$\ln y = \beta_1 x + \beta_0 + \varepsilon$$

$$\ln y = \beta_1 \ln x + \beta_0 + \varepsilon$$

β_1 表示:

X	Y
x (单位值)	y (β_1 单位值)
lnx (100%)	y (β_1 单位值)
x (单位值)	lny (100 β_1 %)
lnx (1%)	lny (β_1 %)

固定效应 FE

首先，固定效应是一种控制变量，可以影响解释变量，也可以影响被解释变量。

其次，固定效应是有固定不变特征的影响因素（控制变量），包括可观测的和不可观测的。

最后，固定效应有两个维度，一个是空间，一个是时间。

Fixed Effects

For a regression model:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \gamma_i + \delta_t + \varepsilon_{it}$$

Averaging i respect to t :

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} + \gamma_i + \bar{\varepsilon}_i$$

Thus

$$\gamma_i = \bar{y}_i - \beta_0 - \beta_1 \bar{x}_{1i} - \beta_2 \bar{x}_{2i} - \bar{\varepsilon}_i$$

Rearranging the original model:

$$\check{y}_{it} \equiv y_{it} - \bar{y}_i = \beta_0 + \beta_1 \check{x}_{1it} + \beta_2 \check{x}_{2it} + \delta_t + \check{\varepsilon}_{it}$$

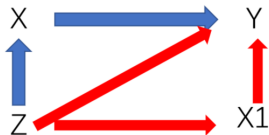
That's “How to control the individual FE”, and the same to time FE.

固定效应

几个类型：个体固定效应、时间固定效应、省份固定效应、行业固定效应以及交叉固定效应等。

根据独一无二的固定不变的特征来分组，比如个体固定效应就是每个个体不随时间变化的独特特征对被解释变量的影响，省份固定效应就是每个省份不随时间变化的独特特征对被解释变量的影响，不同的分组可以形成不同的固定效应。

工具变量：解决内生性的办法



工具变量（Z）的要求：只能通过X影响Y

- Z能影响X（Z药对X疾病有效，不能是面粉、白开水）
- Z不能通过其他方式影响Y
 - Z不能直接影响Y（若Z的副作用是失眠、嗜睡，不是好工具变量）
 - Z不能通过其他变量影响Y（若Z会使人头疼，则不是好工具变量）

两阶段最小二乘法 (2SLS)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

x_2 内生，工具变量为 z

- 适用条件：扰动项不存在自相关或异方差
- 第一阶段：reg x_2 on x_1 z ，得到 $\hat{x}_2$
- 第二阶段：reg y on x_1 $\hat{x}_2$
- 不要自己进行两次手动回归（会导致残差不正确）
 - ivregress 2sls y x_1 ($x_2 = z_1 z_2$)
 - ivreg2 y x_1 ($x_2 = z_1 z_2$)

本讲你将学到:

```
1  reg
2  ivreg
3  xtreg
4  reghdfe
5  ivreghdfe
6  est
7  outreg2
8  .....
```

一些更为全面的参考资料

- ▶ 固定效应：《固定效应：目前看到解释的最清楚的帖子，救命!》
<https://mp.weixin.qq.com/s/5hxLz7XJ30Z9Jmo9SGEWYA>
- ▶ 《内生性专题 | 工具变量法和 2SLS——ivregress2 命令》
<https://mp.weixin.qq.com/s/HCNdnkO0P8Wb0cHnGINWZg>
- ▶
- ▶ 《Stata：关于 reghdfe 命令常见问题解答》
<https://mp.weixin.qq.com/s/fFLCRphoJSgQEMGEPpqfWw>
- ▶ 《Stata 结果输出：outreg2 命令详解》
<https://mp.weixin.qq.com/s/jcnMO2IMiqsDg0SRcKzEMg>

DO!