统 计 学 Statistics

陈 灯 塔

Econ, XJTUCC

CDT.WISE@G

2021 年 5 月 15 日

内容梗概

线性关系

1. 两个经济变量间，如何定义线性关系？
2. 相关系数到底测度什么内容？(即经济意义是什么)

变量之间的关系

变量之间在数量上的依存关系 (interdependence, relationship)

1. 确定性关系：给定 X 取值，Y 完全确定
   - 给定 X 的值，Y 有唯一值
   - 通常用函数表示，例如 $Y = a + bX$，$Y = X^2$ (X 为标准正态时，Y 为 χ^2 分布)
2. 非确定性关系: 收入水平 (y) 与受教育程度 (x) 之间的关系。给定 X 取值, 只能知道 Y 的分布
   - 给定 X 的值，Y 有多个值
   - 如果两者不独立，给定 X 取值发生改变时，Y 的分布会相应变化
   - 可能存在函数关系，例如 $Y = a + bX + \epsilon$, $\epsilon \sim \mathrm{N}(0, \sigma^2)$ 且 $\mathrm{cov}(X, \epsilon) = 0$
   - 两个变量独立时，不存在函数关系，例如 $(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, 0)$

我们关心的，都能写成函数关系，差别在于有没有干扰项！用概率工具

相关分析

定性分析

p184 图 7-1 相关关系类型示意图

相关表: 将具有相关关系的原始数据, 按某一顺序平行排列在一张表上, 以观察它们之间的相互关系　　p185 表 7-1

相关图: 也称为分布图或散点图，相关图所反映的变量之间的相关关系的方向和程度，更为直观　　p186 图 7-2

相关系数

The Pearson product-moment correlation coefficient (a line of best fit through a dataset of two variables)

$$\rho = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} \to r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

$$= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} = \frac{\frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n} \sum x_i^2 - \bar{x}^2} \sqrt{\frac{1}{n} \sum y_i^2 - \bar{y}^2}}$$

计算机中大数计算相对不精确

- $-1 \leqslant r \leqslant 1$
- $r = \pm 1$ 时完全线性相关, $Y = a + bX$
- $r = 0$ 表示没有线性关系 (线性拟合效果最差). 可以存在非线性关系，例如 $Y = X^2$ 其中 X 服从标准正态

相关系数

Excel 计算: CORREL, PEARSON; 分析工具库中的“相关系数”工具

经验划分:

- 弱相关 $0 < |r| < 0.3$,
- 低度相关 $0.3 \leqslant |r| < 0.5$,
- 中度相关 $0.5 \leqslant |r| < 0.8$,
- 高度相关 $0.8 \leqslant |r| < 1$ (工程里 0.9 可能还是弱相关)

相关系数进行检验: $\mathbb{H}_0 : \rho = 0$, 统计量

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$$

相关系数

究竟度量的什么？

- 教科书在讨论相关关系时，文字定义指的往往是依赖关系 (非确定性关系), 而数学分析时采用的是相关系数，偷换概念了
- p187 相关系数是指线性相关条件下反映两个变量之间相关反向和密切程度的统计指标。
  问题：什么是线性相关？

非确定性关系中，如何定义线性关系？

1. $Y = a + bX + \epsilon$, $\mathrm{E}(\epsilon \mid X) = 0$
2. $Y = a + bX + \epsilon$, $\mathrm{E}(\epsilon X) = \mathrm{E}(\epsilon) = 0$ (或者 $\mathrm{cov}(\epsilon X) = \mathrm{E}(\epsilon) = 0$)
3. $Y = a + bX + \epsilon$, $\mathrm{E}(\epsilon) = 0$ 但允许 $\mathrm{E}(\epsilon X) \neq 0$ (经济计量学中考虑的内生问题)

相关系数用来测度线性近似 (拟合) 的好坏 [第三种定义下 OLS 是有偏的]

函数关系和相关关系

函数关系和相关关系

课本上把变量的关系分为函数关系和相关关系：该分类相互交叉，且并集非全集

- 函数关系: 可以相关，可以不相关。假设 X 为标准正态，$Y = X^2$ 与 X 不相关；$Y = X^3$ 则 $\rho_{XY} = \frac{3}{\sqrt{15}} = 0.774$
- 相关关系: 可以存在函数关系，例如前面的 $Y = X^3$. 也可能不存在函数关系, 例如 $(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 其中 $\rho \neq 0$
- 两正态变量独立且联合二元正态分布: 没有函数关系，也没有相关关系
- 概率独立，可能有函数关系: 课本 p143 $\bar{x}$ 与 $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n-1}$ 独立

相关系数

确定性关系中，也可以计算相关系数 (回顾前一幻灯片)

- 假设 X 为标准正态，$Y = X^2$ 与 X 不相关
- 假设 X 为标准正态，若 $Y = X^3$ 则 $\rho_{XY} = \frac{3}{\sqrt{15}} = 0.774$

相关系数到底度量什么?

不管是非确定性关系 (例如 $Y = X^3 + \epsilon$, $(X, \epsilon) \sim \mathrm{N}(\mu_X, \mu_\epsilon, \sigma_X^2, \sigma_\epsilon^2, 0)$), 还是确定性关系 (例如 $Y = X^3$, $X \sim \mathrm{N}(0, 1)$), 无论线性还是非线性，相关系数都是用来测度线性近似 (OLS 估计) 的好坏: for which minimizes

$$\min_{a,b \in \mathbb{R}} \mathrm{E}([Y - (a + bX)]^2)$$

with mean square error

$$\mathrm{var}(Y)(1 - \rho_{XY}^2)$$

相关关系与因果关系

因果关系

美国印第安纳州的地区教会想要筹款兴建新教堂，提出教堂能洁净人们的心灵，减少犯罪，降低监狱服刑人数的口号

- 为了增进民众参与的热诚和信心，教会的神父收集了近 15 年的教堂数与在监狱服刑的人数进行统计分析。
- 最近 15 年教堂数与监狱服刑人数呈显著的正相关。那么是否可以由此得出，教堂建得越多，就可能带来更多的犯罪呢？

经过统计学家和教会神父深入讨论，发现监狱服刑人数的增加和教堂数的增加都与人口的增加有关。教堂数的增加并非监狱服刑人数增加的原因。

总结

1. 变量之间在数量上的依存关系 (理解)
2. 相关系数的定义和计算 (掌握)

Keyword: 确定性关系, 非确定性关系, 相关表, 相关图, 线性关系, 线性近似

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- $*$ Y227-stu.xlsx 的 Sheet1 (纸版或者电子版都不需要提交)
- Read
  1. Textbook: S7.4
  2. Reference book: corresponding chapters

2 最小二乘法

- OLS 估计
- 回归分析
- 总结

回归分析

The term **regression** was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean)

- 回退、退化 (Galton) → 回归 (Yule and Pearson) [退化容易理解成一代不如一代 → 稳定平衡]
- Before 1970, it sometimes took up to 24 hours to receive the result from one regression
- The most common form of regression analysis is linear regression (by ordinary least squares)
- Quantile regression: estimates the conditional median (or other quantiles)
- Orthogonal regression: minimizes the orthogonal (perpendicular) distances from the data to the fit line

回归分析步骤

回归分析是根据已知变量估计未知变量的一种统计方法 (回归可简单地定义为在给定 X 值的条件下 Y 值分布的均值)

1. 建立模型
2. 系数估计和检验
3. 修正和改进
4. 政策分析和或预测

OLS 估计

基本假设

课本 p191 的假设有缺陷: 假设 x 不是随机变量，就没有必要再假设 $\operatorname{cov}(\epsilon_i, x_i) = 0$

通常假设为 (也可以参考 wiki: regression analysis)

1. 条件均值为线性函数:
$$\mathrm{E}(y \mid x) = \beta_0 + \beta_1 x \iff y = \beta_0 + \beta_1 x + \epsilon, \mathrm{E}(\epsilon \mid x) = 0$$
2. 外生性: 样本点 $\mathrm{E}(\epsilon_i \mid x_i) = 0$
3. 条件同方差且不存在自相关: $\operatorname{var}(\epsilon_i \mid \mathbf{x}) = \sigma_\epsilon^2$, $\operatorname{cov}(\epsilon_i, \epsilon_j \mid \mathbf{x}) = 0$. $\mathbf{x} = [x_1, x_2, \cdots, x_K]$. 一元回归中 $\mathbf{x} = [1, x]$ (有常数项)
4. 不存在完全共线性: x_i 不能为同一值，至少有两个取值, 即存在 $i \neq j$, 使得 $x_i \neq x_j$ (一元回归中，表现为 x_i 不能为常数)

样本回归方程
$$y = a + bx + e$$

OLS 估计

最小化 $Q = \sum e_i^2 = \sum (a + bx_i - y_i)^2$

关于 a 求偏导数

$$0 = \frac{\partial Q}{\partial a} = 2\sum (a + bx_i - y_i)$$

得

$$a = \bar{y} - b\bar{x}$$

代入关于 b 的偏导

$$0 = \frac{\partial Q}{\partial b} = 2\sum (a + bx_i - y_i)x_i$$

得

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

拟合方程 $\hat{y} = a + bx$

回归分析

概念

- 截距 (intercept) β_0
- 斜率 (slope) β_1: 当自变量每增加或减少一个单位时，因变量所增加或减少的平均数量
- 误差 (error term) ϵ
- 残差 (residual term) e: $\sum e_i = \sum y_i - \hat{y}_i = 0$

相关分析与回归分析：通常没有必要单独进行相关分析

- $r = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2}} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (y_i-\bar{y})^2}}$ 计算工作量比
  $b = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$ 更多
- 回归分析是相关分析的拓展：政策分析和或预测
- ✗利用一元线性回归方程进行回归分析的前提: 所分析的两个变量之间必须存在相关关系，且相关程度是高度相关 (遗漏变量与 X 独立; 当 ρ_{XY} 很小时，可能 β 很大)

例子

p192–3 E7-3 用 Excel 计算 `E7-3.xlsx`

- 用数据分析计算
- 数学公式自己算
- 函数 `INTERCEPT` 和 `SLOPE`

检验和预测

- 可决系数 (Coefficient of determination, 也称为判断系数): $R^2 = r^2_{x\hat{y}}$ (一元回归 $R^2 = r^2_{xy}$)
- 回归标准差
- 检验
  - F 检验: $\beta_1 = 0$ (全体斜率系数, 一元回归时等同于 t 检验)
  - t 检验: $\beta_1 = 0$ (逐个斜率系数)
- 预测 $\mathrm{E}(y|x) \to \hat{y} = a + bx$

  区间估计: 不确定性来源于干扰的不确定性和系数估计的不确定性。给定某个 x

$$r = y - \hat{y} = \beta_0 + \beta_1 x + \epsilon - (a + bx) = [(\beta_0 - a) + (\beta_1 - b)x] + \epsilon$$

  - Mean Response: 只考虑系数估计的不确定性 (取 $\epsilon = 0$), $\hat{y}_0 \pm t_{\frac{\alpha}{2}} \cdot s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$
  - Individual Response: 考虑系数估计和干扰项的影响

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}} \cdot s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} + 1}$$

多元回归

- 理论分析需要使用矩阵

$$y_t = \mathbf{x}_t'\mathbf{b} + e_t \qquad \mathrm{E}\left(e_t \,|\, \mathbf{x}_t\right) = 0$$

其中

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}_{K\times 1} \qquad \mathbf{x}_t = \begin{bmatrix} x_{t1} \\ x_{t2} \\ \vdots \\ x_{tK} \end{bmatrix}_{K\times 1} \qquad t = 1, 2, \cdots, T$$

- 计算实现需要借助计算机

总结

1. 一元线性回归的基本假设 (理解)
2. 一元线性回归的参数估计，即最小二乘法 OLS 估计 (掌握)
3. 检验和预测，多元回归 (了解)

Keyword: 条件均值函数, 外生性, OLS 估计, 可决系数, 多元回归

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- $*$ Y227-stu.xlsx 的 Sheet2 (电子版自己保留, 纸版抄录 Sheet1 Sheet2 数值结果提交, 图形结果无需提交)
- Read
  1. Textbook: S4.1–2
  2. Reference book: corresponding chapters