

统计学基础

咕咕嘎嘎

目录

第 1 章 数理统计基础	1
第 2 章 线性模型	2
2.1 一元线性模型	2
2.1.1 回归直线	2
2.1.2 最小二乘估计	3
2.1.3 最小二乘估计的性质	4
2.1.4 极大似然估计	4
2.1.5 一元线性回归模型的检验	5
2.1.6 一元线性回归模型参数的区间估计	7
2.1.7 预测与控制	7
第 3 章 时间序列分析	8
第 4 章 统计模拟	9
第 5 章 抽样方法	10

第 1 章 数理统计基础

第2章 线性模型

内容提要

□ 一元线性模型

□ 多元线性模型

我们希望从数据上找到两个现象之间的关系，比如说物理学中我们寻找质量不变时加速度和力的关系到底是怎样的？我们应该用什么函数模型来描述现象之间的关系？为了解决这些问题，我们把得到的数据画到坐标纸上，试图用一条线来经过尽可能多的点——我们称这种行为为“拟合”。这里就蕴含了线性回归的思想：尝试用一个“回归方程”来靠近尽可能多的数据点，并使这个方程与所有点的“距离”尽可能地小，从而得到一个数学式的解释。我们在线性模型中就要解决这种问题：我们要建立什么样的模型，在什么样的标准下才能最小化“距离”之和？这个模型和它的参数又有什么性质？这个模型真实可靠、有现实意义吗？下面我们从一元线性模型开始阐述这个课题。

2.1 一元线性模型

一元回归分析研究一个变量（因变量 y ）对一个解释变量（自变量 x ）的依从关系。

2.1.1 回归直线

我们假设，在建模合理的情况下（变量种类、数量选择合理，它们的关系也处理得当）数据在回归直线旁两侧随机波动。也就是说：如果我们设回归直线为 $f(x) = E[Y|X]$ 的话，那么

$$y = E[Y|X] + \mu.$$

这里的条件期望我们称为回归函数，它是在已知 x 时对 y 在均方误差意义下最好的估计量。关于条件期望的详细内容，见 An&P 条件期望部分内容。需要注意的是，这里的 x 虽然被称为自变量，但是我们把它看作固定的常数。事实上，它们就是我们收集到的数据，自然是已知的。而这里随机性的来源是 μ ，它是一个随机变量，因此 y 也是一个随机变量。 μ 就是数据在回归直线两侧的波动，我们称其为离差，或随机误差项。它是无法观测的。离差代表了样本点和回归方程的距离。

在线性回归中，我们取回归函数 $E[Y|X]$ 为线性函数——这里的“线性”是相对于系数 β_i 而非自变量 x_i 而言的。一元线性回归中，我们取回归函数

$$E[Y|X] = \beta_0 + \beta_1 x.$$

于是模型就变为

$$y_i = \beta_0 + \beta_1 x_i + \mu_i.$$

我们称这个式为总体回归函数，我们在散点图中画的拟合直线并不是总体回归函数，因为总体回归函数总是未知的。注意，这里的 β_0, β_1 都是未知但是固定的数，我们要估计这两个参数。

所谓的“拟合”就是通过估计模型的参数试图预测两个现象之间的数量关系。我们手上拿着两种数据：自变量 x_i 和因变量 y_i ，现在就要通过估计的模型算出因变量的估计 \hat{y}_i ，这要通过估计模型中的两个参数 β_0, β_1 来实现，这时

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

这就是样本回归函数。我们用样本回归函数估计总体回归函数。

在试图用样本回归函数估计了总体回归函数后，我们当然不能期望两者完全一样。也就是存在残差 $\hat{\mu}_i$ ：

$$y_i = \hat{y}_i + \hat{\mu}_i.$$

以上就是用样本回归函数估计总体回归函数的想法。我们现在更细致一些来讨论线性回归模型的假设和性质。

2.1.2 最小二乘估计

经典线性回归模型 (CLRM) 对随机项 μ 做了以下假设, 这些假设使最小二乘估计得出的参数估计具有优良性质:

1. 条件均值为 0:

$$E[\mu_i|X_i] = 0.$$

2. 同方差:

$$V[\mu_i|X_i] = \sigma_\mu^2.$$

3. 无序列相关:

$$\text{cov}(\mu_i, \mu_j|X_i, X_j) = 0, i \neq j.$$

4. 正态假设:

$$\mu_i \sim N(0, \sigma_\mu^2).$$

如何估计模型中的参数 β_0, β_1 ? 我们想画出来的拟合线当然是越靠近数据点越好。但是怎么定义“近”这个概念? 自然联想到距离。但是欧氏距离在计算上不好处理, 因此用估计直线和样本点之间的距离平方和, 也就是残差平方和

$$Q = \sum_{i=1}^n \mu_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

来衡量。我们要最小化 Q , 这只要求导出极值点即可。

证明 分别对 β_0, β_1 求偏导:

$$\begin{aligned} \left. \frac{\partial Q}{\partial \beta_0} \right|_{\beta_0 = \hat{\beta}_0} &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \\ \left. \frac{\partial Q}{\partial \beta_1} \right|_{\beta_1 = \hat{\beta}_1} &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0. \end{aligned}$$

整理上面的方程:


$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} &= \bar{y}, \\ \hat{\beta}_0 \left(\sum_{i=1}^n x_i \right) + \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 \right) &= \sum_{i=1}^n x_i y_i. \end{aligned}$$

于是得到

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{L_{xy}}{L_{xx}}. \end{aligned}$$

其中

$$\begin{aligned} L_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}, \\ L_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n \bar{x}^2. \end{aligned}$$

 **笔记** 实际上求偏导直接得出的方程组说明了残差的两条性质:

1. $\sum_{i=1}^n \mu_i = 0$;
2. $\sum_{i=1}^n \mu_i x_i = 0$.

2.1.3 最小二乘估计的性质

在经典线性回归模型的假设下，最小二乘估计具有很好的性质。

性质

1. 线性性: $\hat{\beta}_0, \hat{\beta}_1$ 都是 y_i 的线性函数，于是 $\hat{\beta}_0, \hat{\beta}_1$ 都服从正态分布；
2. 无偏性: $E[\hat{\beta}_0] = \beta_0, E[\hat{\beta}_1] = \beta_1$ ；
3. 有效性 (Gauss-Markov): 它是所有线性无偏估计量中具有最小方差的估计；

证明 我们证前两条，并顺带得出两个参数的方差，进而得到两个参数的分布。

1. 首先，可以证明 $\hat{\beta}_1$ 可以改写成

$$\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{j=1}^n (x_j - \bar{x})^2},$$

这就把 $\hat{\beta}_1$ 改写成了 y_i 的线性组合。同理我们能把 $\hat{\beta}_0$ 写成

$$\frac{\sum_{i=1}^n [\frac{L_{xx}}{n} - \bar{x}(x_i - \bar{x})] y_i}{L_{xx}}.$$

2. 现在证明无偏性。我们知道 y_i 也服从正态分布 $N(\beta_0 + \beta_1 x_i, \sigma^2)$ ，于是由期望线性性就知道

$$E[\hat{\beta}_1] = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} E[y_i] = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} (\beta_0 + \beta_1 x_i) = \beta_1.$$

其中用了 $\sum_{i=1}^n (x_i - \bar{x}) = 0, \sum_{i=1}^n (x_i - \bar{x}) x_i = \sum_{i=1}^n (x_i - \bar{x})^2$ 。

于是 β_0 就有：

$$E[\beta_0] = E[\bar{y} - \hat{\beta}_1 \bar{x}] = \beta_0 + \beta_1 x_i - \beta_1 x_i = \beta_0.$$

下面我们来看两个参数的方差。由于 μ_i 互相独立，因此也有 y_i 互相独立。于是


$$V[\hat{\beta}_1] = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{L_{xx}} \right)^2 V[y_i] = \frac{\sigma^2}{L_{xx}}.$$

因此 $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$ 。

另外，

$$V[\hat{\beta}_0] = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{L_{xx}} \right)^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}} \right) \sigma^2.$$

因此 $\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}) \sigma^2)$ 。

 **笔记** 我们从上面得出的参数分布可知，对于不同于 $x_i, i = 1, 2, \dots, n$ 的样本点 x_0 来说，它的预测值

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 = \frac{\sum_{i=1}^n ((x_i - \bar{x})(x_0 - \bar{x}) + \frac{L_{xx}}{n}) y_i}{L_{xx}},$$

同样服从正态分布，即

$$\hat{y}_0 \sim N(\beta_0 + \beta_1 x_0, (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}) \sigma^2).$$

我们得出的样本回归线有几条性质：

1. 过点 (\bar{x}, \bar{y}) ；
2. $E[\hat{y}_i] = E[y_i]$ ；
3. 残差估计值和 \hat{y}_i, x_i 都不相关。

2.1.4 极大似然估计

从极大似然估计的角度，我们也可以得出 $\hat{\beta}_0, \hat{\beta}_1$ ，还有 σ^2 的估计。首先写出似然函数，由 $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ ：

$$f(y; x_1, x_2, \dots, x_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right),$$

取对数:

$$\log f = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (y - \beta_0 - \beta_1 x_i)^2}{2\sigma^2},$$

对于 $\hat{\beta}_0, \hat{\beta}_1$, 求最值的步骤和最小二乘法相同。下面我们看 σ^2 的极大似然估计。对 σ^2 求偏导并令其为零:

$$\frac{\partial}{\partial \sigma^2} \log f = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (y - \beta_0 - \beta_1 x_i)^2}{2(\sigma^2)^2} = 0,$$

解得

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_i^2.$$

然而 $\hat{\sigma}^2$ 并不是 σ^2 的无偏估计。我们经常使用的是无偏估计版本:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\mu}_i^2.$$

2.1.5 一元线性回归模型的检验

在拟合之后得到的模型在描述现象之间的关系时是否合适? 例如, 如果我们研究收入和消费时, 发现收入越多消费反而越少, 这时我们就会不禁怀疑做出的模型是否可靠。实际上我们需要一种客观的方法来检验模型是否有实际意义, 也就是说检验自变量是否真的对因变量有影响? 哪些自变量对因变量有影响? 这样的影响是合乎道理的吗? 下面我们就根据假设检验原理导出检验模型有效性的工具。

统计检验分为模型参数显著性检验 (t 检验)、模型显著性检验 (F 检验), 拟合优度检验、相关系数检验。

先看 t 检验。t 检验的对象是模型的单个参数是否显著, 也就是对于总体回归函数

$$y_i = \sum_{j=0}^n \beta_j x_{ij} + \mu_i, \quad x_{i0} = 1,$$

1. 原假设 $H_0: \beta_j = 0$;
2. 对立假设 $H_1: \beta_j \neq 0$.

在一元回归模型中, 我们只考虑 $j = 1$ 的情况。我们知道 $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$, 所以当 H_0 成立时, $\hat{\beta}_1 \sim N(0, \frac{\sigma^2}{L_{xx}})$, 由于方差未知, 因此可以构造检验统计量

$$t = \frac{\hat{\beta}_1 / \sqrt{(\frac{\sigma^2}{L_{xx}})}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}} \sim t(n-2),$$

其中 $\hat{\sigma}^2$ 的定义取 σ^2 的无偏估计。所以上述检验的拒绝域为 (水平为 α)

$$W = \{t: |t| > t_{\alpha/2}(n-2)\}.$$



笔记 正态总体 $N(\mu, \sigma^2)$ 中抽取的样本 $x_i, i = 1, 2, \dots, n$ 满足

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \sim \chi_{n-1}^2,$$

而残差平方和由于在解最小二乘估计方程组时添加了两个条件, 因此自由度为 $n-2$, 也就是

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi_{n-2}^2.$$

我们记残差平方和为 SSE , 则上式变为

$$\frac{SSE}{\sigma^2} \sim \chi_{n-2}^2.$$

然后是 F 检验。F 检验又称模型显著性检验, 它检验的对象是整个模型中的自变量是否至少有一个对因变量有关。所以一元线性回归中的 t 检验和 F 检验实际上是等价的。

1. 原假设 $H_0: \beta_j = 0, j = 1, 2, \dots, n$;
2. 对立假设 $H_1: \neg H_0$.

我们由平方和分解：

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

简记为

$$SST = SSR + SSE \quad (S_T = S_R + S_E).$$

其中 SST 是总离差平方和，它代表了实际数据和回归直线之间的距离，自由度为 $n-1$ ； SSR 是回归平方和，代表了拟合出的模型解释数据波动的能力，也就是预料之内的波动程度，自由度为 1； SSE 是残差平方和，它代表了数据拟合的优良程度，也代表着模型没能解释的数据波动情况，即意料之外的波动程度，自由度为 $n-2$ 。于是我们想看去除自由度的影响后，衡量这个模型能解释多少数据波动情况，就得出了统计量 F ：

$$F = \frac{SSR/1}{SSE/(n-2)}.$$

它代表了回归直线能解释的数据波动比例。

实际上， $\frac{SSR}{\sigma^2} \sim \chi_1^2$ ， $\frac{SSE}{\sigma^2} \sim \chi_{n-2}^2$ ，因此

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-2}.$$

检验的拒绝域为 (水平为 α)

$$W = \{F : F > F_\alpha(1, n-2)\}.$$

第三个检验是相关系数的显著性检验，这并不是一个假设检验，只是算出一个系数来判断两个因素之间是否有相关性、相关性有多强。

我们称

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{L_{xy}}{\sqrt{L_{xx}} \sqrt{L_{yy}}}$$

为简单相关系数，简称相关系数。它衡量了因变量、自变量之间的线性相关程度。由 Cauchy-Schwartz: $|r| \leq 1$ 。

由于

$$r = \frac{L_{xy}}{\sqrt{L_{xx}} \sqrt{L_{yy}}} = \hat{\beta}_1 \sqrt{\frac{L_{yy}}{L_{xx}}},$$

因此所谓线性关系显著就代表了 $\hat{\beta}_1$ 是否显著。

下面给出三种检验的关系：

1. **t 检验和 F 检验：**我们知道，自由度为 1, n 的 F 分布随机变量可以写成

$$F_{1,n} = \frac{\chi_1^2/1}{\chi_n^2/n},$$

而 $\chi_1^2 = (N(0,1))^2$ ，因此还有

$$F_{1,n} = \frac{(N(0,1))^2}{\chi_n^2/n},$$

而自由度为 n 的 t 分布的构成为 $\frac{N(0,1)}{\sqrt{\chi_n^2/n}}$ ，因此

$$F_{1,n} = t_n^2.$$

从而我们知道 **t 检验和 F 检验** 在一元回归中是等价的。

2. **相关系数检验和 F 检验：**实际上 $SST = L_{yy}$ ， $SSR = \hat{\beta}_1^2 L_{xx}$ 。因此对于 **F 检验**，我们有

$$\frac{1}{1 + \frac{n-2}{F}} = \frac{SSR}{SST} = \hat{\beta}_1^2 \frac{L_{xx}}{L_{yy}} = r^2.$$

最后一个检验是拟合优度检验，它比较了不同自变量对因变量的解释程度，也就是解决“哪个指标能更好的解释因变量”这个问题。我们引进“解释程度”变量：

$$R^2 = \frac{SSR}{SST} = (r)^2.$$

这个比例代表了样本偏差中回归偏差的占比，越靠近 1 说明拟合效果越好。

2.1.6 一元线性回归模型参数的区间估计

假设检验和置信区间都应用了枢轴量来构造，因此具有天然的一一对应关系。我们考察 β_0, β_1 的置信区间。我们知道 $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$ ，由于方差未知，我们构造 t 置信区间，枢轴量为

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{L_{xx}}}} \sim t_{n-2},$$

置信区间如下构造：

$$P(|\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{L_{xx}}}}| \leq t_{\alpha/2}(n-2)) = 1 - \alpha,$$

解得 $1 - \alpha$ 置信区间为

$$\hat{\beta}_1 \pm \frac{t_{\alpha/2}(n-2)\hat{\sigma}}{\sqrt{L_{xx}}}.$$

对于 $\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}})\sigma^2)$ 也可同理构造 t 统计量。

2.1.7 预测与控制

单值控制：

1. $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$;
2. $E[\hat{y}_0] = E[y_0] = E[Y|X = x_0]$.

这两条分别代表了对于个体的预测和对于群体均值的预测。

区间预测：我们已经说明了预测值 \hat{y}_0 和实际值 y_0 的分布为

$$\hat{y}_0 \sim N(\beta_0 + \beta_1 x_0, (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}})\sigma^2), y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2),$$

并且 \hat{y}_0, y_0 相互独立：因为 \hat{y}_0 是 $y_i, i = 1, 2, \dots, n$ 的线性函数，而 y_0 和其他数据都无关，因此 $y_0 - \hat{y}_0$ 同样服从正态分布

$$N(0, (1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}})\sigma^2).$$

这样我们又可以得到 y_0 水平为 $1 - \alpha$ 的 t 置信区间：

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}})\hat{\sigma}^2} = \hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{1 + h_{00}}\hat{\sigma}.$$


其中 $h_{00} = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}$ 。若选择 $\alpha = 5\%$ ，则置信区间约为 $\hat{y} \pm 2\hat{\sigma}$ 。

同理可得 $\hat{y}_0 - E[y_0]$ 的分布为

$$N(0, (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}})\sigma^2),$$

得到 $E[y_0]$ 水平为 $1 - \alpha$ 的 t 置信区间：

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{h_{00}}\hat{\sigma}.$$

 **笔记** 对比得到，均值预测的置信区间半径小于单值预测置信区间，因为总体更稳定。

第 3 章 时间序列分析

第 4 章 统计模拟

第 5 章 抽样方法