统计方法与机器学习

第二章：一元线性回归分析

倪 蒨

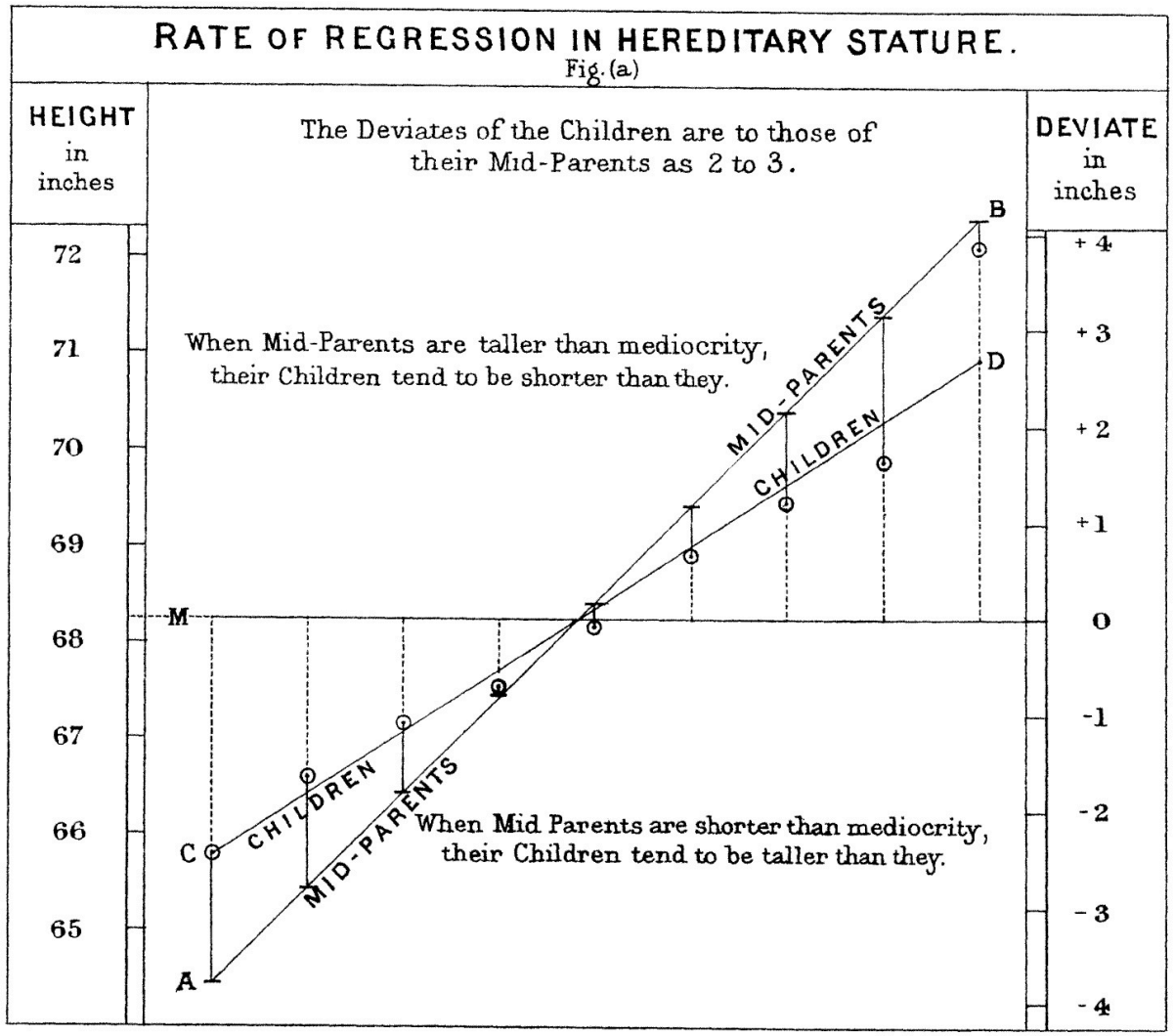
DaSE@ECNU
(lni@dase.ecnu.edu.cn)

目录

目录

线性回归的背景

故事


RATE OF REGRESSION IN HEREDITARY STATURE.
Fig. (a)
HEIGHT in inches
DEVIATE in inches
The Deviates of the Children are to those of their Mid-Parents as 2 to 3.
When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they.
When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.
MID-PARENTS
CHILDREN
CHILDREN
MID-PARENTS
A
B
C
D
M
72
71
70
69
68
67
66
65
+4
+3
+2
+1
0
-1
-2
-3
-4

线性回归的背景

故事

- 回归分析是由 19 世纪英国统计学家 F·高尔顿（F·Galton）首先提出的。
- 一个有趣的现象：
  - 一群特高个子父母的儿子们在同龄人中在平均的意义上属于高个子；
  - 一群高个子父母的儿子们在同龄人中在平均的意义上属于略高个子；
  - 一群特矮个子父母的儿子们在同龄人中在平均的意义上属于矮个子；
  - 一群矮个子父母的儿子们在同龄人中在平均的意义上属于略矮个子。
- 高尔顿提出了用“**回归**”一词来概括父母平均身高 x 和其儿子的身高 y 之间的关系。

线性回归的背景

故事

- K·皮尔逊收集了 1078 对父子的身高数据；
- x 表示父亲的身高；
- y 表示成年儿子的身高；
- 建立回归直线方程，即

$$\hat{y} = 33.73 + 0.516x$$

- 主要结论：
  - 父亲身高每增加 1 个单位，其儿子的身高平均增加 0.516 个单位。
  - 高个子父亲有生高个子儿子的趋势，但是一群高个子父辈的儿子们的平均高度要低于父辈的平均高度。
  - 矮个子父辈的儿子们虽为矮个子，但是其平均身高要比父辈高一些。

线性回归的背景

概述

- 在实际问题中，感兴趣的变量 y 与易于获得的变量 x 之间存在紧密关联，但又不由变量 x 而唯一确定的，这种关系通常称为**统计关系**。
- 若变量 y 与 x 间有统计关系，
  - 通常称 y 为因变量或响应变量；
  - x 为自变量或解释变量，这里 x 在机器学习方法中也会被称为特征。
- 给定 x 的取值后，y 的取值是无法唯一确定的。
- 于是，我们可以将 y 认为一个随机变量，并需要通过概率分布来对它进行描述，而我们常常关心的是这个概率分布的数字特征，如：期望和方差。

线性回归的背景

概述

- 给定 x 时，称 y 的条件数学期望为 y 关于 x 的（均值）回归函数，即

$$f(x) = E(y|x)$$

- 注意到，$f(x)$ 不仅是 x 的一个确定性的函数，并且从平均意义上刻画了变量 y 与 x 间统计关系的规律。
- 而如何确定这个确定性的函数 f 是回归问题中最为核心的问题。
  - 线性回归模型可看作将这个函数 f 取为 x 的一个线性函数的形式，如 $f(x) = \beta_0 + \beta_1 x$；
  - 神经网络模型可看作将这个函数 f 取为 x 的一个非线性函数的形式，如 $f(x) = \max(0, \beta_0 + \beta_1 x)$；
  - 深度学习模型可理解为这个函数 f 取为 x 的多个非线性函数的复合形式。

目录

一元线性回归模型

概述

- 一元线性回归模型为

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

其中，β_0, β_1 为两个未知参数，常称为回归系数，而 ε 是随机误差。

- 与数学模型

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

在理解上是不同的，主要的差异在于是否引入了**随机误差项** ε：

  - 在数学模型中，两个变量之间的关系是确定性的；
  - 在统计模型中，两个变量之间的关系是不确定的。

一元线性回归模型

概述

- 一元线性回归模型为

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

  其中有两个部分：
  - 确定性的部分是 $\beta_0 + \beta_1 x$；
  - 随机性的部分是 ε。
- 随机误差项用来概括由于人们认识以及其他客观原因的局限而没有考虑的种种偶然因素。
- 例如，
  - 相同年龄的孩子的识字量不同；
  - 相同生产条件下的产品质量存在差异。

一元线性回归模型

概述

- 随机误差 ε 是无法被观测；
- 但通常假定 ε 满足

$$\begin{cases} E(\varepsilon) = 0 \\ \mathrm{Var}(\varepsilon) = \sigma^2 < \infty \end{cases}$$

其中，$E(\varepsilon)$ 表示 ε 的数学期望；$\mathrm{Var}(\varepsilon)$ 表示 ε 的方差。

一元线性回归模型

概述

- 一元线性回归模型为

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

- 由于随机误差项 ε 存在，y 也是随机变量；
- 关于 x 求条件期望，即

$$E(y|x) = \beta_0 + \beta_1 x.$$

- 注意到 $E(y|x)$ 是关于 x 的一个函数，表示用 x 的信息刻画因变量 y，作为 y 的” 预测”。
- 称 $E(y|x) = \beta_0 + \beta_1 x$ 为**回归方程**。

一元线性回归模型

概述

- 在回归方程

$$E(y|x) = \beta_0 + \beta_1 x.$$

中，由于回归系数 β_0, β_1 均是未知的，因此，我们需要通过所观测到的数据 $(x_i, y_i), i = 1, 2, \cdots, n$ 进行估计。

- 一般假定数据 $\{(x_i, y_i)\}_{i=1}^n$ 符合线性回归模型及其假设，即

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n, \\ E(\varepsilon_i) = 0 \quad \text{和} \quad \mathrm{Var}(\varepsilon_i) = \sigma^2 \end{cases}$$

并假定 n 组数据是独立观测的，即假定 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 是独立同分布的随机变量。

一元线性回归模型

概述

- 因为，我们认为 $x_1, x_2 \cdots, x_n$ 均是确定性变量，是可以精确测量和控制的，
- 所以，$y_1, y_2, \cdots, y_n$ 的期望与方差分别为

$$E(y_i) = \beta_0 + \beta_1 x_i \quad \text{和} \quad \mathrm{Var}(y_i) = \sigma^2, i = 1, 2, \cdots, n.$$

- 这表明，随机变量 $y_1, y_2, \cdots, y_n$
  - 服从不同的分布，方差相等，而期望不等；
  - 但，相互独立。

一元线性回归模型

概述

- 一元线性回归模型的两个“版本”:
- “模型版”

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

- “数据版”

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \cdots, n.$$

- 差异在于前者强调的是模型，后者侧重的是数据。

一元线性回归模型

任务：预测与参数估计

- 回归分析最为常见任务之一是通过 n 组样本观测值 $(x_i, y_i), i = 1, 2, \cdots, n$，对一个新的个体进行预测。
- 具体来说，如果 x_0 已知，那么 $\beta_0 + \beta_1 x_0$ 是 y_0 的一个合理的预测值。
- 而回归系数 β_0 和 β_1 都是未知的常数，因此，我们需要估计在回归模型中的未知参数。
- 之后，我们会介绍两种估计方法。

一元线性回归模型

任务：预测与参数估计

- 一般会用 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别表示 β_0 和 β_1 的估计值。
- y 关于 x 的一元线性**经验**回归方程为

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

- 其图形称为经验回归直线。
  - $\hat{\beta}_0$ 表示经验回归直线的截距；
  - $\hat{\beta}_1$ 表示经验回归直线的斜率。
- 给定 $x = x_0$ 后，称

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

为回归值。有时，也称为拟合值或者预测值。

目录

最小二乘估计

概述

- 对于每一个样本观测值 (x_i, y_i)，定义偏差为观测值 y_i 与其回归值 $E(y_i|x_i)$ 的差异为

$$y_i - E(y_i|x_i) = y_i - \beta_0 - \beta_1 x_i.$$

- 偏差平方和为

$$\begin{aligned} Q(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - E(y_i))^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 . \end{aligned}$$

- 称通过最小化偏差平方和 $Q(\beta_0, \beta_1)$ 而得到的参数估计方法，为最小二乘估计（Least Squares Estimation）。

最小二乘估计

概述

- 最小二乘估计定义为

$$\begin{aligned}\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \arg\min_{\beta_0,\beta_1} Q(\beta_0, \beta_1) \\ &= \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2\end{aligned}$$

- 这本质上是一个求极值问题。
- 因为 Q 是关于 $\hat{\beta}_0, \hat{\beta}_1$ 的非负二次函数，所以其最小值总是存在的。

最小二乘估计

解法

- 要求 $Q(\beta_0, \beta_1)$ 的最小值，令其一阶导数为零，即

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \end{cases}$$

- 经整理后，得到正规方程组

$$\begin{cases} n\beta_0 + n\bar{x}\beta_1 = n\bar{y} \\ n\bar{x}\beta_0 + \sum\limits_{i=1}^{n} x_i^2 \beta_1 = \sum\limits_{i=1}^{n} x_i y_i \end{cases}$$

最小二乘估计

解法

- 于是，β_0, β_1 的最小二乘估计为

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \end{cases}$$

- 其中，

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

分别为 $x_1, x_2, \cdots, x_n$ 和 $y_1, y_2, \cdots, y_n$ 的样本均值。

最小二乘估计

一种简单的记号

- 如果记

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

- 那么，最小二乘估计简写为

$$
\begin{cases}
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \\
\hat{\beta}_1 = l_{xx}^{-1}l_{xy}.
\end{cases}
$$

最小二乘估计

说明

- 根据一阶导等于零，所求的 $\hat{\beta}_0, \hat{\beta}_1$ 实际上是 $Q(\beta_0, \beta_1)$ 的稳定点。
- 但是否为最小值点，仍需要根据其二阶导在 $(\hat{\beta}_0, \hat{\beta}_1)$ 上的表现来判断是否为最小值点。
- 对 $Q(\beta_0, \beta_1)$ 求二阶偏导，我们有

$$\begin{aligned}
\left| \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix} \right| &= 4n \sum_{i=1}^{n} x_i^2 - 4n^2(\bar{x})^2 \\
&= 4n \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&> 0.
\end{aligned}$$

最大似然估计

回顾

- **最大似然估计**是依赖于总体的概率函数 $f(x;\theta)$ 以及样本所提供的信息求未知参数估计。
- 当总体 X 为连续随机变量时，其密度函数为

$$\{f(x;\theta), \theta \in \Theta\}$$

- 假定总体 X 的一个独立同分布的样本为 $x_1, x_2, \cdots, x_n$。参数的似然函数为

$$L(\theta; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

最大似然估计

回顾

- 最大似然估计指的是在参数空间 Θ 中选取随机样本 $(X_1, X_2, \cdots, X_n)$ 落在点 $(x_1, x_2, \cdots, x_n)$ 附近最大概率的 $\hat{\theta}$ 为未知参数 θ 的估计值，即 $\hat{\theta}$ 应满足

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x_1, x_2, \cdots, x_n).$$

最大似然估计

分布假定

- 在一元线性回归模型中，最常见的假定为 ε 服从正态分布，即

$$\varepsilon \sim N(0, \sigma^2)$$

- 从数据的角度来看，由于 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 都是与 ε 独立同分布的随机变量，因而有

$$\varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2), \quad i = 1, 2, \cdots, n$$

- 在 ε_i 服从正态分布的假定下，y_i 也服从正态分布，即

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, 2, \cdots, n.$$

最大似然估计

解法

- y_i 的密度函数为

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\}.$$

- 因为 $y_1, y_2, \cdots, y_n$ 的密度函数的形式是不尽相同的, 我们用 $f_i(y_i)$ 代替 $f(y_i)$ 更为合适。
- 似然函数为

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f_i(y_i) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\} \end{aligned}$$

最大似然估计

解法

- 易知，L 的最大值点与 $\ln L$ 的最大值点是相同的。
- 于是，对数似然函数为

$$\ln L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

最大似然估计

解法：β_0, β_1

- 易知，L 的最大值点与 $\ln L$ 的最大值点是相同的。
- 于是，对数似然函数为

$$\ln L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

- 我们发现

$$\arg\max_{\beta_0,\beta_1} \ln L(\beta_0, \beta_1, \sigma^2) \Leftrightarrow \arg\min_{\beta_0,\beta_1} Q(\beta_0, \beta_1)$$

- 回归系数 β_0, β_1 的最大似然估计和最小二乘估计的形式是一致的。

最大似然估计

解法：σ^2

- 我们可以对 $\ln L(\beta_0, \beta_1, \sigma^2)$ 关于 σ^2 求导，并令一阶导为零，即

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 = 0$$

- σ^2 的最大似然估计为

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{ML}} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2\end{aligned}$$

最大似然估计

说明

- $\hat{\sigma}^2_{\mathrm{ML}}$ 是 σ^2 的有偏估计。
- 在实际应用中，更为常用的是 σ^2 的无偏估计，即

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}\Big(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i)\Big)^2 .\end{aligned}$$

- 最大似然估计是在 $\varepsilon_i \sim N(0,\sigma^2)$ 的正态分布假设下求得的，而最小二乘估计则对分布假设没有要求；
- $y_1, y_2, \cdots, y_n$ 是独立的正态分布样本，而不是同分布的，但这并不妨碍最大似然方法的应用。

最大似然估计

定理 1

如果 $y_1, y_2, \cdots, y_n$ 是相互独立的且 y_i 是正态分布随机变量，即

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2),$$

那么

- $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right), \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right)$;
- $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}}\sigma^2$;

最大似然估计

证明

- 可以将 $\hat{\beta}_1$ 写为以下的形式

$$\begin{aligned}
\hat{\beta}_1 &= l_{xx}^{-1} l_{xy} = l_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
&= l_{xx}^{-1} \left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i - \sum_{i=1}^{n} (x_i - \bar{x}) \bar{y} \right) \\
&= l_{xx}^{-1} \left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i \right) \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{l_{xx}} y_i
\end{aligned}$$

第三个等式是因为 $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$;

最大似然估计

证明

- 可以将 $\hat{\beta}_0$ 写为以下的形式

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

- 因为自变量 $x_1, x_2, \cdots, x_n$ 是确定性的，且 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 均可以看作 $y_1, y_2, \cdots, y_n$ 的线性组合。
- 已知 $y_1, y_2, \cdots, y_n$ 是相互独立的正态随机变量，那么 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 均服从正态分布。

最大似然估计

证明

- 可以将 $\hat{\beta}_0$ 写为以下的形式

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

- 因为自变量 $x_1, x_2, \cdots, x_n$ 是确定性的，且 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 均可以看作 $y_1, y_2, \cdots, y_n$ 的线性组合。
- 已知 $y_1, y_2, \cdots, y_n$ 是相互独立的正态随机变量，那么 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 均服从正态分布。

最大似然估计

证明

- 接下来，我们需要考虑这两个估计的均值与方差，从而进一步确定分布。
- 一方面，考虑 $\hat{\beta}_1$ 的期望与方差，即

$$
\begin{aligned}
E(\hat{\beta}_1) &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{l_{xx}} E(y_i) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{l_{xx}} (\beta_0 + \beta_1 x_i) \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{l_{xx}} \beta_0 + \sum_{i=1}^{n} \frac{x_i(x_i - \bar{x})}{l_{xx}} \beta_1 \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{l_{xx}} \beta_0 + \sum_{i=1}^{n} \frac{(x_i - \bar{x})x_i}{l_{xx}} \beta_1 - \sum_{i=1}^{n} \frac{\bar{x}(x_i - \bar{x})}{l_{xx}} \beta_1 \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{l_{xx}} \beta_0 + \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{l_{xx}} \beta_1 = \beta_1 \\
\mathrm{Var}(\hat{\beta}_1) &= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{l_{xx}} \right)^2 \mathrm{Var}(y_i) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(l_{xx})^2} \sigma^2 = \frac{\sigma^2}{l_{xx}}
\end{aligned}
$$

最大似然估计

证明

- 另一方面，考虑 $\hat{\beta}_0$ 的期望与方差，即

$$
\begin{aligned}
E(\hat{\beta}_0) &= E(\bar{y}) - E(\hat{\beta}_1)\bar{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\
&= \beta_0 \\
\mathrm{Var}(\hat{\beta}_0) &= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}}\right)^2 \mathrm{Var}(y_i) \\
&= \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n^2} - \frac{2(x_i - \bar{x})\bar{x}}{n l_{xx}} + \frac{(x_i - \bar{x})^2 \bar{x}^2}{l_{xx}^2}\right) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right)
\end{aligned}
$$

最大似然估计

证明

- 因为 $y_1, y_2, \ldots, y_n$ 相互独立，所以

$$
\begin{aligned}
\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \mathrm{Cov}\left(\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}}\right) y_i, \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{l_{xx}} y_i\right) \\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}}\right) \frac{(x_i - \bar{x})}{l_{xx}} \sigma^2 \\
&= -\frac{\bar{x}}{l_{xx}} \sigma^2.
\end{aligned}
$$

最大似然估计

说明

根据上述定理，我们可得到以下一些有用的推论。

- $\hat{\beta}_0, \hat{\beta}_1$ 分别是 β_0, β_1 的无偏估计；
- 除 $\bar{x} = 0$ 外，$\hat{\beta}_0$ 与 $\hat{\beta}_1$ 是相关的；
- 为了提高 $\hat{\beta}_0, \hat{\beta}_1$ 的估计精度（即降低它们的方差）就要求样本量 n 增加，或使得 l_{xx} 增大，即要求 $x_1, x_2, \cdots, x_n$ 比较分散。

目录

回归方程的显著性检验

概述

- 建立回归方程的目的是寻找 y 的均值随 x 变化的规律，即找出回归方程 $E(y) = \beta_0 + \beta_1 x$。
- 什么叫回归方程有意义呢?

回归方程的显著性检验

概述

- 建立回归方程的目的是寻找 y 的均值随 x 变化的规律，即找出回归方程 $E(y) = \beta_0 + \beta_1 x$。
- 什么叫回归方程有意义呢?
- 如果 $\beta_1 = 0$，那么不管 x 如何变化，$E(y)$ 不随 x 的变化作线性变化，那么称回归方程不显著。
- 如果 $\beta_1 \neq 0$，那么当 x 变化时，$E(y)$ 随 x 的变化作线性变化，那么称回归方程是显著的。

回归方程的显著性检验

概述

- 对回归方程是否有意义作判断就是要判断回归直线的斜率是否为零。
- 提出合适的检验问题为

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

- 如果我们得到的结论是拒绝 H_0，那么我们认为回归方程是显著的。
- 接下来，我们会介绍三种检验方法。

F 检验

定义

- 运用**方差分析**的思想；
- 令回归值为

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

和残差为

$$e_i = y_i - \hat{y}_i$$

- 偏差平方和为

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = l_{yy}.$$

其中，$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$。

F 检验

定义

引起各 y_i 不同的原因主要有两类因素：

- 其一是 H_0 可能不真，即 $\beta_1 \neq 0$，从而 $E(y) = \beta_0 + \beta_1 x$ 随 x 的变化而变化，即在每一个 x 的观测值处的回归值不同，定义回归平方和为

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

- 其二是其他一切因素。在得到回归值之后，y 的观测值与回归值之间还有差异，定义为残差平方和

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

F 检验

定义

- 一元线性回归场合下的平方和分解式

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

即

$$SS_T = SS_R + SS_E$$

F 检验

定理 2

设

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

其中 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 相互独立，且

$$E(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2, i = 1, 2, \cdots, n$$

我们有

$$\begin{aligned} E(SS_R) &= \sigma^2 + \beta_1^2 l_{xx} \\ E(SS_E) &= (n-2)\sigma^2 \end{aligned}$$

F 检验

证明

由于回归平方和 SS_R 可写为

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

从而

$$\begin{aligned} E(SS_R) &= E\left(\hat{\beta}_1^2\right) l_{xx} = \left(\mathrm{Var}(\hat{\beta}_1) + (E(\hat{\beta}_1))^2\right) l_{xx} \\ &= \left(\frac{\sigma^2}{l_{xx}} + \beta_1^2\right) l_{xx} \\ &= \sigma^2 + \beta_1^2 l_{xx}. \end{aligned}$$

证明

而残差平方和为

$$
\begin{aligned}
SS_E &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(\beta_0+\beta_1x_i+\varepsilon_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2\\
&= \sum_{i=1}^{n}((\beta_0-\hat{\beta}_0)+(\beta_1-\hat{\beta}_1)x_i+\varepsilon_i)^2\\
&= \sum_{i=1}^{n}\Big((\beta_0-\hat{\beta}_0)^2+(\beta_1-\hat{\beta}_1)^2x_i^2+\varepsilon_i^2+2(\beta_0-\hat{\beta}_0)(\beta_1-\hat{\beta}_1)\\
&\quad +2(\beta_0-\hat{\beta}_0)\varepsilon_i+2(\beta_1-\hat{\beta}_1)x_i\varepsilon_i\Big).
\end{aligned}
$$

证明

于是有

$$\begin{aligned} E(SS_E) &= n\mathrm{Var}(\hat{\beta}_0) + \mathrm{Var}(\hat{\beta}_1)\sum_{i=1}^{n} x_i^2 + n\mathrm{Var}(\varepsilon_i) + 2n\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_0) \\ &\quad -2\sum_{i=1}^{n} E(\hat{\beta}_0\varepsilon_i) - 2\sum_{i=1}^{n} x_i E(\hat{\beta}_1\varepsilon_i). \end{aligned}$$

由于 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 可写成 $y_1, \cdots, y_n$ 的线性组合.

$$\hat{\beta}_0 = \sum_i \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}}\right) y_i; \quad \hat{\beta}_1 = \sum_i \frac{(x_i - \bar{x})}{l_{xx}} y_i.$$

证明

利用 y_i 之间是独立的, 有

$$
\begin{aligned}
E(\hat{\beta}_0\varepsilon_i) &= E\left(\varepsilon_i \sum_j \left(\frac{1}{n} - \frac{(x_j-\bar{x})\bar{x}}{l_{xx}}\right) y_j\right) \\
&= E\left(\varepsilon_i \sum_j \left(\frac{1}{n} - \frac{(x_j-\bar{x})\bar{x}}{l_{xx}}\right) (\beta_0 + \beta_1 x_j + \varepsilon_j)\right) \\
&= E\left(\varepsilon_i^2 \left(\frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right)\right) \\
&= \left(\frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right)\sigma^2.
\end{aligned}
$$

证明

$$
\begin{aligned}
E(\hat{\beta}_1 \varepsilon_i) &= E\left(\varepsilon_i \sum_j \frac{(x_j - \bar{x})}{l_{xx}} y_j\right) \\
&= E\left(\varepsilon_i \sum_j \frac{(x_j - \bar{x})}{l_{xx}} (\beta_0 + \beta_1 x_j + \varepsilon_j)\right) \\
&= \frac{(x_i - \bar{x})}{l_{xx}} \sigma^2.
\end{aligned}
$$

证明

由此,

$$\begin{aligned}
\sum_{i=1}^{n} E(\hat{\beta}_0 \varepsilon_i) &= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}} \right) \cdot \sigma^2 = \sigma^2, \\
\sum_{i=1}^{n} E(\hat{\beta}_1 \varepsilon_i) &= \sum_{i=1}^{n} x_i \frac{(x_i - \bar{x})}{l_{xx}} \sigma^2 = \sigma^2.
\end{aligned}$$

于是

$$\begin{aligned}
E(SS_E) &= n \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} \right) \sigma^2 + \frac{\sigma^2}{l_{xx}} \cdot \sum_i x_i^2 + n\sigma^2 - 2n \cdot \frac{\bar{x}^2}{l_{xx}} \sigma^2 - 2\sigma^2 - 2\sigma^2 \\
&= \sigma^2 \left(1 + \frac{\sum x_i^2}{l_{xx}} - \frac{n\bar{x}^2}{l_{xx}} + n - 4 \right) \\
&= (n-2)\sigma^2.
\end{aligned}$$

F 检验

定理 3

设 $y_1, y_2, \cdots, y_n$ 相互独立，且

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), i = 1, 2, \cdots, n$$

则有

- $SS_E/\sigma^2 \sim \chi^2(n-2)$;
- 若 H_0 成立，则有 $SS_R/\sigma^2 \sim \chi^2(1)$;
- SS_R 与 SS_E，$\bar{y}$ 独立。

证明

首先，我们可以构造一个正交矩阵 $\boldsymbol{A}$，形如

$$\boldsymbol{A}=\{a_{ij}\}_{n\times n}=\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p} \\ \vdots & \vdots & & \vdots \\ a_{n-2,1} & a_{n-2,2} & \cdots & a_{n-2,p} \\ \frac{x_1-\bar{x}}{\sqrt{l_{xx}}} & \frac{x_2-\bar{x}}{\sqrt{l_{xx}}} & \cdots & \frac{x_n-\bar{x}}{\sqrt{l_{xx}}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \end{pmatrix}.$$

根据矩阵的正交性质 $\boldsymbol{A}\boldsymbol{A}'=\boldsymbol{I}$，其中，$\boldsymbol{A}'$ 表示矩阵 $\boldsymbol{A}$ 的转置矩阵，$\boldsymbol{I}$ 是单位阵。于是，$\boldsymbol{A}$ 满足

$$\sum_k a_{i,k}a_{j,k}=0,\quad 1\le i<j\le n-2;$$

$$\frac{1}{\sqrt{n}}\sum_j a_{i,j}=0;$$

$$\sum_j a_{i,j}\frac{x_j-\bar{x}}{\sqrt{l_{xx}}}=0;$$

F 检验

证明

- 值得注意的是，矩阵 $\boldsymbol{A}$ 总共有 $n(n-2)$ 个未知参数，而上述有 $3(n-2)+\binom{n-2}{2}=\frac{(n-2)(n+3)}{2}$ 个方程。
- 只要 $n \geq 3$，未知参数个数不少于方程个数，因此，正交矩阵 $\boldsymbol{A}$ 一定是存在的.

令 $\boldsymbol{z}=\boldsymbol{A}\boldsymbol{y}$，其中 $\boldsymbol{z}=(z_1, z_2, \cdots, z_n)'$ 满足

$$
\begin{aligned}
z_i &= \sum_j a_{ij} y_j, \quad i=1, \cdots, n-2; \\
z_{n-1} &= \frac{\sum_j (x_j-\bar{x}) y_j}{\sqrt{l_{xx}}} = \frac{\sum_j (x_j-\bar{x})(y_j-\bar{y})}{\sqrt{l_{xx}}} = \frac{l_{xy}}{\sqrt{l_{xx}}} = \sqrt{l_{xx}} \cdot \frac{l_{xy}}{l_{xx}} = \sqrt{l_{xx}} \hat{\beta}_1; \\
z_n &= \frac{1}{\sqrt{n}} \sum_j y_j = \sqrt{n} \bar{y}.
\end{aligned}
$$

F 检验

证明

那么，$\boldsymbol{z}$ 仍服从 n 维正态分布，且其均值与方差分别为

$$\begin{aligned}
E(z_i) &= E(\sum_j a_{ij}y_j) = \sum_j a_{ij}(\beta_0 + \beta_1 x_j) \\
&= \beta_0 \sum_j a_{ij} + \beta_1 \sum_j a_{ij}x_j = 0, \quad i = 1, \cdots, n-2; \\
E(z_{n-1}) &= \sqrt{l_{xx}} \cdot \beta_1; \\
E(z_n) &= \sqrt{n}(\beta_0 + \beta_1 \hat{x}); \\
\mathrm{Var}(\boldsymbol{z}) &= \mathrm{Var}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{y})\boldsymbol{A}' = \boldsymbol{A}\sigma^2\boldsymbol{I}_n\boldsymbol{A}^2 = \sigma^2\boldsymbol{I}_n.
\end{aligned}$$

证明

我们可以得到以下结论：

- $z_1, z_2, \cdots, z_n$ 相互独立；
- 前 $n-2$ 个分量 $x_1, x_2, \cdots, x_{n-2}$ 是独立同分布的，且分布为 $N(0, \sigma^2)$；
- z_{n-1} 的分布为 $N\left(\sqrt{l_{xx}}\beta_1, \sigma^2\right)$；
- z_n 的分布为 $N\left(\sqrt{n}(\beta_0 + \beta_1\bar{x}), \sigma^2\right)$。

F 检验

证明

因为

$$\begin{aligned}
\sum_{i=1}^{n} z_i^2 &= \boldsymbol{z}'\boldsymbol{z} = \boldsymbol{y}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{y} = \sum_i y_i^2 = SS_T + n\bar{y}^2, \\
z_{n-1} &= \sqrt{l_{xx}}\hat{\beta}_1 = \sqrt{SS_R}, \\
z_n &= \sqrt{n}\bar{y}.
\end{aligned}$$

证明

所以，我们可以整理为

$$SS_T + n\bar{y}^2 = \sum_{i=1}^{n-2} z_i^2 + SS_R + n\bar{y}^2$$

即

$$SS_T = \sum_{i=1}^{n-2} z_i^2 + SS_R$$

因此，

- SS_E 的分布为

$$SS_E = SS_T - SS_R = \sum_{i=1}^{n-2} z_i^2 \sim \chi^2(n-2).$$

F 检验

证明

- 在 $\beta_1 = 0$ 时, 因为

$$z_{n-1} \sim N(0, \sigma^2),$$

即 z_{n-1}/σ 是标准正态分布的随机变量，所以

$$\frac{SS_R}{\sigma^2} = \left(\frac{z_{n-1}}{\sigma}\right)^2 \sim \chi^2(1)$$

- 因为 SS_E 与前 $n-2$ 个 x_i 有关，SS_R 仅与 z_{n-1} 有关，而 $\bar{y}$ 仅与 z_n 有关，因此 SS_R 与 SS_E，$\bar{y}$ 相互独立。

- 因为 $\hat{\beta}_1$ 仅与 SS_R 有关，所以 $\hat{\beta}_1$ 与 SS_E，$\bar{y}$ 相互独立。

F 检验

检验统计量

- 考虑构造形如

$$F_0 = \frac{SS_R}{SS_E/(n-2)}$$

  的检验统计量来检验。
- 在 $\beta_1 = 0$ 时，$F \sim F(1, n-2)$。对于给定的显著性水平 α，其拒绝域为

$$F_0 \geq F_{1-\alpha}(1, n-2),$$

  其中，$F_\alpha(\mathrm{df}_1, \mathrm{df}_2)$ 表示自由度分别为 df_1 和 df_2 的 F 分布的 α 分位数。
- 这个检验称为 F 检验。

F 检验

检验统计量

表: 一元回归分析的方差分析表

来源	平方和	自由度	均方	F 值
回归	SS_R	1	$MS_R = SS_R$	$F_0 = \frac{SS_R}{SS_E/(n-2)}$
误差	SS_E	$n-2$	$MS_E = \frac{SS_E}{n-2}$	
总和	SS_T	$n-1$		

t 检验

检验统计量

- 由于

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right), \quad \frac{SS_E}{\sigma^2} \sim \chi^2(n-2),$$

且 $\hat{\beta}_1$ 相互独立，因此在 H_0 为真时，有

$$t_0 = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}} \sim t(n-2)$$

其中 $\hat{\sigma} = \sqrt{SS_E/(n-2)}$。

- 对于给定的显著性水平 α，拒绝域为

$$W = \left\{|t_0| > t_{1-\alpha/2}(n-2)\right\}.$$

相关系数的检验

概述

- 由于一元线性回归方差能否反映两个随机变量 x 与 y 间的线性相关关系时，它的显著性检验还可以通过对二维总体相关系数 ρ 的检验进行。
- 其假设为

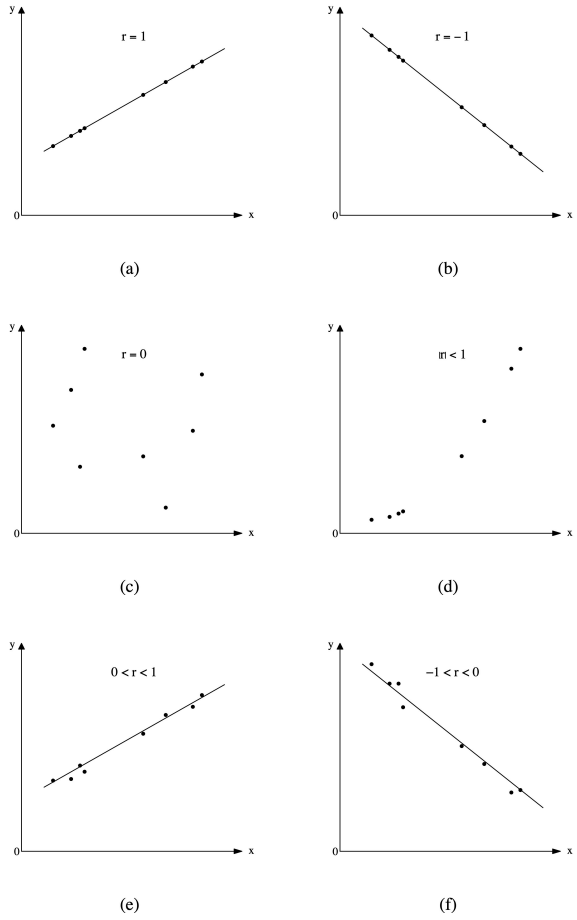
$$H_0 : \rho = 0, \quad \text{vs} \quad H_1 : \rho \neq 0.$$

- $\{(x_i, y_i) : i = 1, \cdots, n\}$ 可看作容量为 n 的二维样本。
- 所用的检验统计量为样本相关系数

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

相关系数的检验

概述



相关系数的检验

检验统计量

- H_0 为真时，$|r|$ 应较小。
- 当 $|r|$ 较大时，应拒绝原假设。
- 因此，拒绝域为 $\{|r| \geq c\}$，其中，临界值 c 可由 H_0 成立时样本相关系数的分布确定，该分布与自由度 $n-2$ 有关。
- 对给定的显著性水平 α，由

$$P(W) = P(|r| \geq c) = \alpha$$

知, 临界值 c 应是 $H_0 : \rho = 0$ 成立下 r 的分布的 $1-\alpha/2$ 分位数，故记为 $c = r_{1-\alpha/2}(n-2)$。

- 如何得到这个临界值呢?

三种检验之间的关系

考虑 t 检验统计量与 F 检验统计量的关系

$$t_0^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}} \right)^2 = \frac{\hat{\beta}_1^2 l_{xx}}{\sqrt{SS_E/(n-2)}} = \frac{SS_R}{SS_E/(n-2)} = F_0$$

其中，第三个等式成立是因为回归平方和 SS_R 与 $\hat{\beta}_1$ 之间存在如下关系：

$$\begin{aligned} SS_R &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} [\bar{y} + \hat{\beta}_1 (x_i - \bar{x}) - \bar{y}]^2 = \sum_{i=1}^{n} [\hat{\beta}_1 (x_i - \bar{x})]^2 = \hat{\beta}_1^2 l_{xx} \end{aligned}$$

因此，F 检验与 t 检验是等价的。

三种检验之间的关系

其次，考虑 F 检验统计量与样本相关系数 r 的关系。我们有

$$
\begin{aligned}
r^2 &= \left(\hat{\beta}_1\sqrt{\frac{l_{xx}}{l_{yy}}}\right)^2 = \hat{\beta}_1^2\frac{l_{xx}}{l_{yy}} = \frac{SS_R}{SS_T} = \frac{SS_R}{SS_R+SS_E} \\
&= \frac{SS_R/(SS_E/(n-2))}{SS_R/(SS_E/(n-2))+(n-2)} = \frac{F_0}{F_0+(n-2)}
\end{aligned}
$$

这表明了 $|r|$ 是 F_0 的严格单调增函数，因此可以从 F 分布的 $1-\alpha$ 分位数 $F_{1-\alpha}(1,n-2)$ 得到相关系数检验所需要确定的临界值 $r_{1-\alpha/2}(n-2)$，即

$$
r_{1-\alpha/2}(n-2) = \sqrt{\frac{F_{1-\alpha}(1,n-2)}{F_{1-\alpha}(1,n-2)+(n-2)}}
$$

三种检验之间的关系

这里 r^2 也常常作为回归分析中一项重要的指标。定义样本决定系数为回归平方和与总偏差平方和之比，即

$$r^2 = \frac{SS_R}{SS_T}$$

样本决定系数 r^2 是一个回归直线与样本观测值拟合优度的相对指标，反映因变量的波动中能用自变量解释的比例，r^2 的取值在 0 到 1 之间。r^2 越接近 1，拟合优度越好。

三种检验之间的关系

说明

- 三种检验方法在一元线性回归模型下是等价的；
- 但在多元线性回归场合，经推广 F 检验仍可用，另两个检验就无法使用了；
- 如果无法拒绝原假设，则可以认为回归方程是不显著的，导致这种情况的可能原因如下：
  - 误差与正态假设严重背离；
  - Y 与 X 无关；
  - Y 与 X 虽然相关，但不是线性关系；
  - Y 与 X 以外的因素有更密切的关系。

估计与预测

概述

当 $x = x_0$ 时，我们关心的是

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0,$$

注意到，y_0 是本身一个随机变量。我们需要明确一下估计和预测这两个问题在定义上的差异。

- 当 $x = x_0$ 时，想要确定其均值 $E(y_0) = \beta_0 + \beta_1 x_0$，由于 $E(y_0)$ 是一个常数，可以记为 μ_0，而非随机变量，因此，这是一个**估计问题**。对于估计而言，我们可以进一步讨论如何确定 μ_0 的点估计与区间估计。

估计与预测

概述

当 $x = x_0$ 时，我们关心的是

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0,$$

注意到，y_0 是本身一个随机变量。我们需要明确一下估计和预测这两个问题在定义上的差异。

- 当 $x = x_0$ 时，y_0 本身的范围是什么？由于 y_0 是随机变量，如果仅用一个数来表示一个随机变量，那么通过采用均值，但数字特征无法刻画 y_0 的分布信息。在确定一个随机变量而言，我们可以构造一个区间，使得 y_0 落在这个区间的概率为 $1-\alpha$，即确定一个常数 δ，使得 $P(|y_0 - \hat{y}_0| \leq \delta) = 1 - \alpha$，称区间 $[\hat{y}_0 - \delta, \hat{y}_0 + \delta]$ 为 y_0 的概率为 $1-\alpha$ 的预测区间，这是一个**预测问题**。

关于 $E(y_0)$ 的估计

点估计

- 在 $x = x_0$ 时，我们需要考虑 $E(y_0) = \beta_0 + \beta_1 x_0$，一个直观的估计为

$$\hat{E}(y_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

通常记为 $\hat{y}_0$，表示在 $x = x_0$ 时响应变量的估计值。

关于 $E(y_0)$ 的估计

定理 4

如果 $y_1, y_2, \cdots, y_n$ 是相互独立的且 y_i 是正态分布随机变量，即 $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, 那么，对给定的 x_0，

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}\right)\sigma^2\right).$$

关于 $E(y_0)$ 的估计

证明

根据定理 1 可知，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别都是 $y_1, y_2, ..., y_n$ 的线性组合。在给定 x_0 时，$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 也是 $y_1, y_2, ..., y_n$ 的线性组合，因此 $\hat{y}_0$ 也服从正态分布，其均值和方差为

$$
\begin{aligned}
E(\hat{y}_0) &= E(\hat{\beta}_0) + E(\hat{\beta}_1)x_0 = \beta_0 + \beta_1 x_0 \\
\mathrm{Var}(\hat{y}_0) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \mathrm{Var}(\hat{\beta}_0) + \mathrm{Var}(\hat{\beta}_1)x_0^2 + 2\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)x_0 \\
&= \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right) \cdot \sigma^2 + \frac{x_0^2}{l_{xx}} \cdot \sigma^2 - \frac{2\bar{x}x_0}{l_{xx}} \cdot \sigma^2 \\
&= \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}\right) \sigma^2
\end{aligned}
$$

所以，$\hat{y}_0 \sim N\left(\beta_0 + \beta_1 x_0, \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{xx}}\right)\sigma^2\right)$。

关于 $E(y_0)$ 的估计

区间估计

- $\hat{y}_0$ 的分布，即

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}\right)\sigma^2\right).$$

- 在构造区间时，由于 σ^2 时一个未知常数，因此，我们需要用其估计代替，而

$$SS_E/\sigma^2 \sim \chi^2(n-2).$$

- 同时，我们注意到 $\hat{y}_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x})$ 与 SS_E 是相互独立。

关于 $E(y_0)$ 的估计

区间估计

- 于是，我们有

$$\frac{(\hat{y}_0 - E(y_0))/\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})}{l_{xx}}\sigma^2}}{\sqrt{\frac{SS_E}{\sigma^2}/(n-2)}} = \frac{\hat{y}_0 - E(y_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{xx}}}} \sim t(n-2)$$

- 因此，$E(y_0)$ 的置信水平为 $1-\alpha$ 的置信区间为

$$[\hat{y}_0 - \delta_0, \hat{y}_0 + \delta_0],$$

其中，

$$\delta_0 = t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{xx}}}.$$

关于 y_0 的预测

点预测

- 在预测 y_0 时，考虑点预测时，也就是说，用一个数来刻画一个随机变量，通常我们还是取这个随机变量的均值，于是，y_0 的点预测也是 $\hat{y}_0$。
- 但是，由于 y_0 是一个连续的随机变量，恰好取到一个点的概率为零，因此，在实际应用中，对 y_0 进行区间预测更为合理。

关于 y_0 的预测

区间预测

- 事实上，$y_0 = E(y_0) + \varepsilon_0$，因为通常假定 $\varepsilon_0 \sim N(0, \sigma^2)$，所以 y_0 的最有可能取值仍为 $\hat{y}_0$。
- 于是，我们可以使用以 $\hat{y}_0$ 为中心的一个区间

$$[\hat{y}_0 - \delta_1, \hat{y}_0 + \delta_1]$$

作为 y_0 的取值范围。如何确定 δ_1 的值是需要进一步讨论的。

关于 y_0 的预测

区间预测

- 一方面，我们知道

$$y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$$

其点预测 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 是 $y_1, y_2, \cdots, y_n$ 的线性组合，且服从正态分布，即

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

又因为 y_0 与 $\hat{y}_0$ 独立，所以

$$y_0 - \hat{y}_0 \sim N\left(0, \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}\right)\sigma^2\right).$$

关于 y_0 的预测

区间预测

- 而另一方面，因为 $(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2)$，而且 $y_0, \hat{y}_0, \hat{\sigma}^2$ 相互独立。所以，我们有

$$\frac{y_0 - \hat{y}_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim t(n-2).$$

- 因此，预测区间为

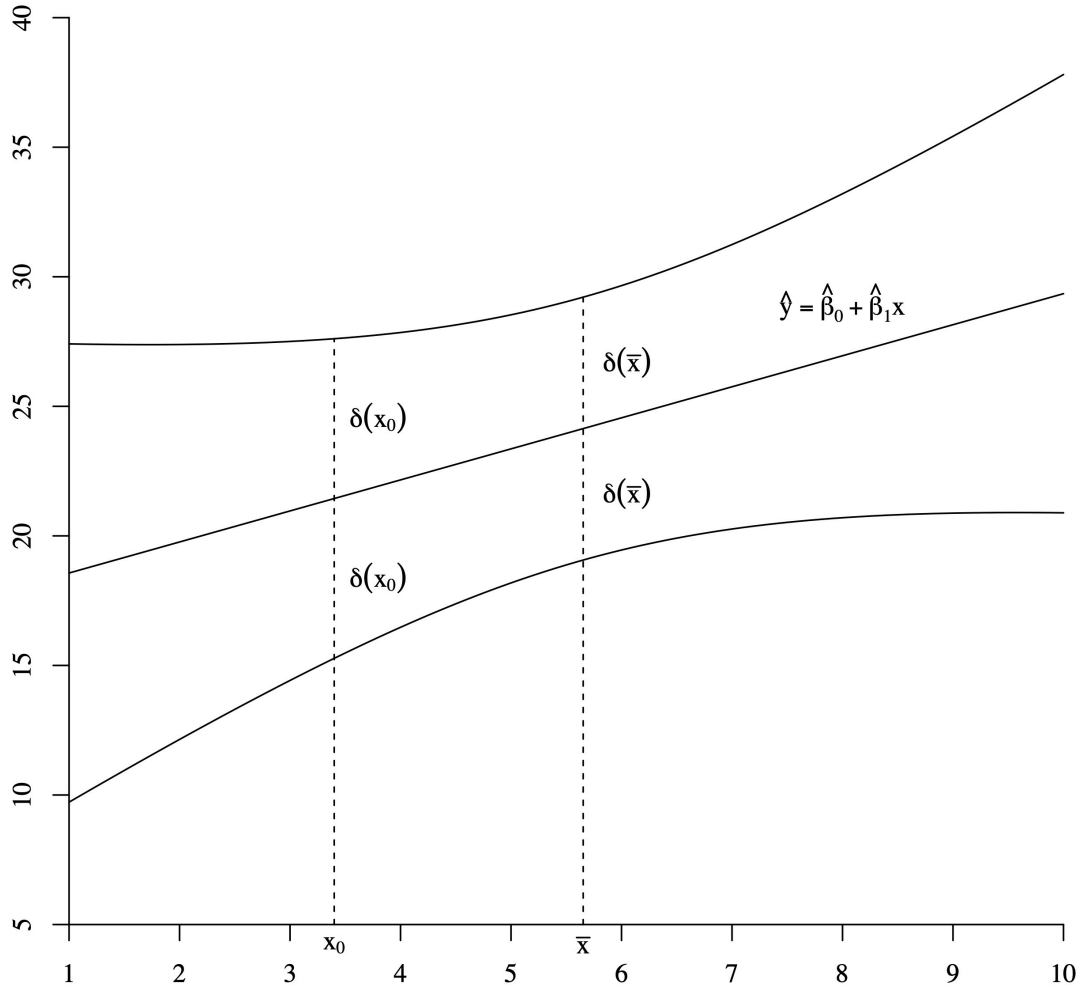
$$[\hat{y}_0 - \delta_1, \hat{y}_0 + \delta_1],$$

其中，δ_1 为

$$\delta_1 = \delta(x_0) = t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}$$

关于 y_0 的预测

说明


40
35
30
25
20
15
10
5
1
2
3
4
5
6
7
8
9
10
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\delta(\bar{x})$
$\delta(x_0)$
$\delta(\bar{x})$
$\delta(x_0)$
x_0
$\bar{x}$

估计与预测

说明

- 区间的形式为

$$[\hat{y}_0 - \delta, \hat{y}_0 + \delta],$$

- 估计区间中 δ 为

$$\delta = t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}.$$

- 预测区间中 δ 为

$$\delta = t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}.$$