回归分析

- 回归分析的概念
- 一元线性回归分析
- 多元线性回归分析

1　回归分析的概念

在现实问题中，处于同一个过程中的一些变量，往往是相互依赖和相互制约的，它们之间的相互关系大致可分为两种：

（1）确定性关系——函数关系；

（2）非确定性关系——相关关系；

相关关系表现为这些变量之间有一定的依赖关系，但这种关系并不完全确定，它们之间的关系不能精确地用函数表示出来，这些变量中至少有一个是随机变量。

correlation(相关)—how two data sets change together
causation(因果)—whether or not one data set influences another

相关关系举例

- 在其它条件基本相同时，某农作物的亩产量 Y 与施肥量 X 之间有一定的关系，但施肥量相同，亩产量却不一定相同。亩产量是一个随机变量。

- 人的血压 Y 与年龄 X 之间有一定的依赖关系，一般来说，年龄越大，血压越高，但年龄相同的两个人的血压不一定相等。血压是一个随机变量。

农作物的亩产量与施肥量、血压与年龄之间的这种关系称为相关关系，在这些变量中，施肥量、年龄是可控变量，亩产量、血压是不可控变量。一般在讨论相关关系问题中，可控变量称为自变量，不可控变量称为因变量。

例 小麦的亩产量记为 Y, 它与水 (x_1) 、肥料 (x_2) 、土质 (x_3) 、麦种 (x_4) 、栽培技术 (x_5) 及管理措施 (x_6) 等因素有关. 由于观测或试验中总存在随机因素的影响, 即使 $x_1, x_2, \cdots, x_6$ 相对固定, 小麦的亩产量也不完全相同, 因此将 Y 与 $x_1, x_2, \cdots, x_6$ 的**相关关系**分为两部分来研究,

$$Y = f(x_1, x_2, \cdots, x_6) + \varepsilon$$

其中 $f(x_1, x_2, \cdots, x_6)$ 表示 6 个可控因素 $x_1, x_2, \cdots, x_6$ 与亩产量 Y 的确定关系, f 是确定性函数, 表示非随机部分; ε 表示随机因素对亩产量 Y 的影响, 一般把 ε 看成数学期望 $E(\varepsilon) = 0$ 的随机变量. 于是, Y 是一个随机变量, 它是可观测的, 其数学期望 $E(Y) = f(x_1, x_2, \cdots, x_6)$.

为数学处理方便起见, 可以近似地把f当作线性函数,

$$f(x_1, x_2, \cdots, x_6) = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6.$$

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6 + \varepsilon$$

这样 Y 是关于 $x_1, x_2, \cdots, x_6$ 的线性函数, 其中 $\beta_0, \beta_1, \cdots, \beta_6$ 是未知参数.

数理统计的一个任务是定量地研究这种非确定性的关系，包括通过观察和试验数据去判断变量之间有无关系，对其关系大小进行估计、推断和预测等.

研究相关关系时，一般可以分为:

- 随机变量与随机变量之间的相关关系
- **随机变量与普通变量之间的相关关系**

这两种情况的假设不同，推导过程也不相同，但某些方法和结论有类似之处. 我们只讨论后一种情况.

回归(regrssion)分析是研究相关关系的一种重要的数理统计方法.

- 只有两个变量的回归分析，称为**一元回归分析(simple regression)**
- 超过两个变量时，称为**多元回归分析(multiple regression)**
- 变量之间成线性关系时, 称为**线性回归(linear regression)**
- 变量之间具有非线性关系时, 称为**非线性回归(non-linear regression)**

回归(regression)

19 世纪, 英国生物学家兼统计学家**高尔顿 (Galton)**发表了论文《身高遗传中的平庸回归》，研究了父与子身高的遗传问题，提出：

- 通常父代身材高大的, 其子代身材也高大;
- 父代身材矮小的, 其子代身材亦矮小;
- 子代的平均高度有**向中心回归**的趋势, 使得一段时间内人的身高相对稳定.

后来, 英国著名统计学家**皮尔逊**观察了 1078 对父子, 用 x 表示父亲身高, y 表示成年儿子的身高, 将 (x, y) 点在直角坐标系中,发现这 1078 个点基本在一条直线附近, 并求出了该直线的方程 (单位: 英寸, 1 英寸 = 2.54 cm):

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

这表明:

- 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.
- 高个子父辈生的儿子平均身高也高, 但子辈的身高间的差距低于父辈间的身高差距 (为 0.516 倍).

之后回归分析的思想渗透到了数理统计的其他分支中. 随着计算机的发展, 各种统计软件包的出现, 回归分析的应用就越来越广泛.

2 一元线性回归

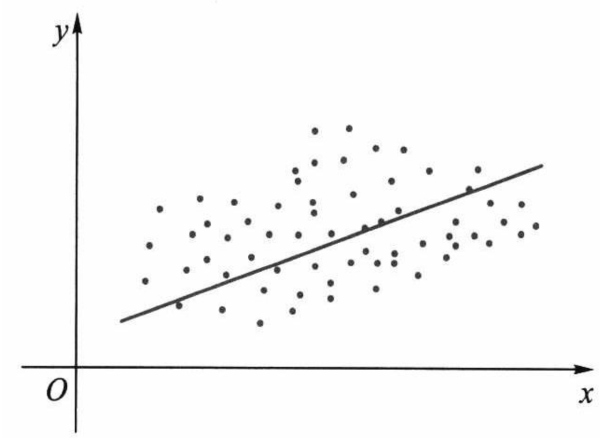
一、一元正态线性回归模型

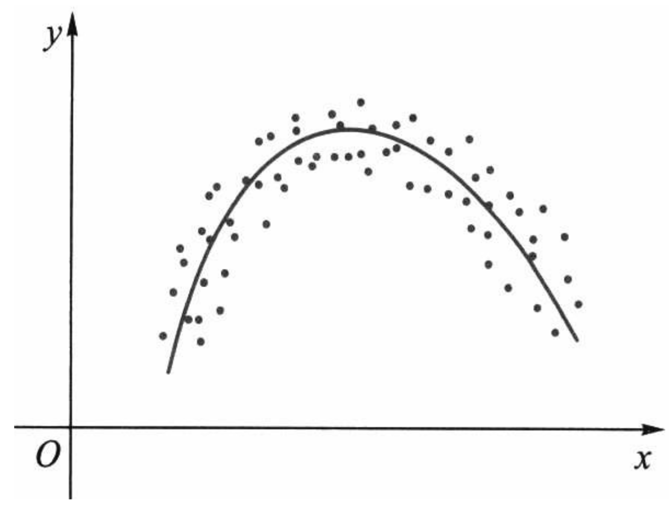
设随机变量Y，对于X的每一个值, Y均有自己的分布 .若$E(Y)$存在，则它一定是X的函数，记为$E(Y) = f(X)$，其值可通过样本进行估计.

对于X的一组值x_i $(i=1,\dots,n)$，作独立试验，对Y得出n个观测结果y_i $(i=1,\dots,n)$，即有n次独立观察，得样本观测值：(x_1,y_1) , (x_2,y_2) ,… , (x_n,y_n)

问题: 如何利用这些样本观测值来估计$f(X)$.

首先要推测$f(X)$的形式，一般可以作出散点图，从中可粗略看出Y与X的关系.





若Y和X之间大体上呈现线性关系，可假定

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

其中β_0和 β_1是未知常数, ε 表示其它随机因素的影响.

通常假定 ε 服从正态分布$N(0,\sigma^2)$，即

$$\begin{cases} E(\varepsilon) = 0 \\ D(\varepsilon) = \sigma^2 > 0 \end{cases}$$ 其中σ^2为未知参数.

$Y = \beta_0 + \beta_1 X + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$ (1)称为一元(正态)线性回归模型.

$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

$E(Y) = \beta_0 + \beta_1 X$，用$E(Y)$作为$Y$的估计 $\hat{Y}$

$\hat{Y} = \beta_0 + \beta_1 X$ (2)称为 Y 关于 X 的一元线性回归方程.

对变量X，Y进行n次独立观察，得样本观测值：

$$(x_1,y_1),\ldots,(x_n,y_n) \qquad (3)$$

由此样本得方程组：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,2,\cdots,n \qquad (4)$$

ε_i是第i次观察时的随机误差，是不可观察的随机变量.

由于各次观察独立，有

$$\begin{cases} E(\varepsilon_i) = 0 \\ D(\varepsilon_i) = \sigma^2 > 0 \end{cases}, i = 1,2,\cdots,n \qquad (5)$$

回归分析的任务是利用n组独立观察数据

$(x_1,y_1),\ldots,(x_n,y_n)$来估计$\beta_0$和$\beta_1$, 以估计值 $\widehat{\beta_0}$ 和 $\widehat{\beta_1}$

分别代替(2)式中的β_0和β_1，得回归方程

$$\hat{Y}=\widehat{\beta_0}+\widehat{\beta_1}X \qquad (6)$$

方程(6)的建立依赖于通过观察或试验取得的数据,称其为经验回归方程或经验公式.

$\widehat{\beta_0}$ 和 $\widehat{\beta_1}$称为未知参数 β_0 , β_1 的回归系数.

问题：如何利用n组独立观察数据来估计β_0和β_1 ?

二、最小二乘估计

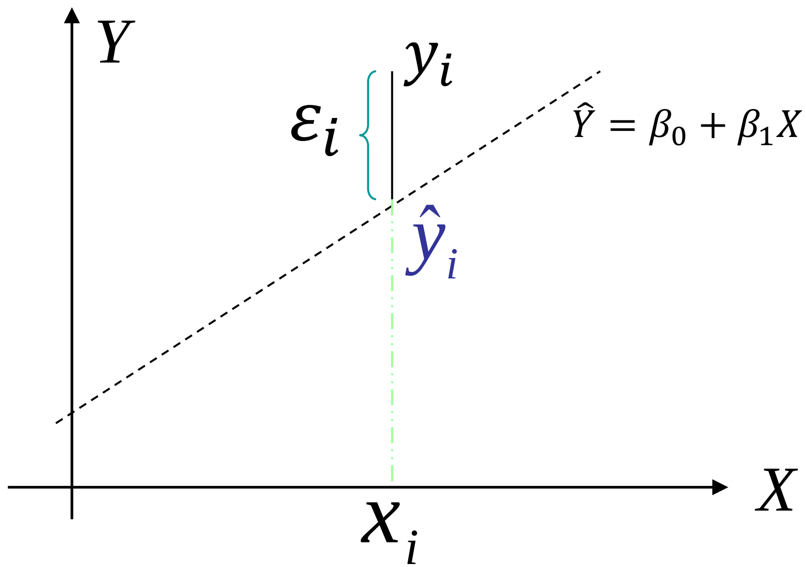
假设为了估计某物体的重量，对它进行了n次称量，因称量有误差，n次称量结果x_1，$x_2,\ldots,x_n$有差异，现在用数$\hat{x}$去估计该物体的重量，则$\hat{x}$与上述n次称量结果的偏差的平方和为:

$$\sum_{i=1}^{n}(x_i-\hat{x})^2$$

一个好的估计$\hat{x}$，应使这个平方和尽可能地小.

估计原则:寻找一个使上述平方和达到最小的$\hat{x}$，作为这个物体重量的估计值。这种方法称为最小二乘法. 用最小二乘法作出的估计叫最小二乘估计.

对(X,Y)作n次观察（试验），得到n对数据，要求找一条直线$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X$，尽可能好地拟合这些数据.



由回归方程，当X取值x_i时，$\hat{y}_i$应取值$\beta_0 + \beta_1 x_i$，而实际观察到的为y_i，这样就形成了偏差

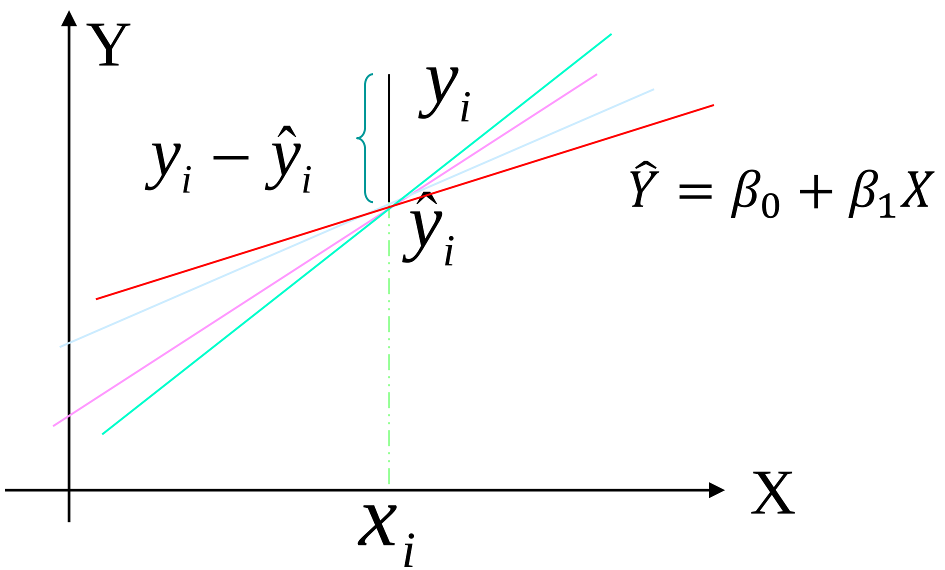
$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i) = y_i - \hat{y}_i$$

依照最小二乘法的思想，提出目标量

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2 \qquad \textbf{(7)}$$

是所有实测值$\mathbf{y}_\mathbf{i}$与回归值 $\hat{y}_i$ 的偏差平方和.

求出β_0 , β_1的估计值$\widehat{\beta_0}$ 和 $\widehat{\beta_1}$, 使偏差平方和$\mathbf{Q}(\beta_0,\beta_1)$达到最小.



采用微积分中求极值的办法，求出使$Q(\beta_0, \beta_1)$达到最小的$\widehat{\beta_0}, \widehat{\beta_1}$.

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

得
$$\left.\begin{aligned} & n \cdot \beta_0 + \left(\sum_{i=1}^{n} x_i\right) \cdot \beta_1 = \sum_{i=1}^{n} y_i \\ & \left(\sum_{i=1}^{n} x_i\right) \cdot \beta_0 + \left(\sum_{i=1}^{n} x_i^2\right) \cdot \beta_1 = \sum_{i=1}^{n} x_i y_i \end{aligned}\right\} \text{正规方程组}$$

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

改写正规方程组 $\begin{cases}\beta_0+\overline{x}\beta_1=\overline{y}\\ n\overline{x}\beta_0+\left(\sum_{i=1}^{n}x_i^2\right)\beta_1=\sum_{i=1}^{n}x_iy_i\end{cases}$

假设$\boldsymbol{x_i}$不全相同，则系数行列式不为**0**，

$$\begin{vmatrix}1 & \overline{x}\\ n\overline{x} & \sum_{i=1}^{n}x_i^2\end{vmatrix}=\sum_{i=1}^{n}x_i^2-n(\overline{x})^2=\sum_{i=1}^{n}(x_i-\overline{x})^2\neq 0$$

方程组有唯一解

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} {x_i}^2 - n(\bar{x})^2} \triangleq \frac{l_{xy}}{l_{xx}}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

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

$$l_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$l_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n(\bar{x})^2$$

最小二乘估计的性质

(1) $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right)$, $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right)$ ($\hat{\beta}_0, \hat{\beta}_1$ 分别是 β_0, β_1 的无偏估计)

(2) $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}}\sigma^2$

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

(4) $\widehat{\beta_0}$ 和 $\widehat{\beta_1}$都是$y_1, y_2, \ldots, y_n$的线性函数, 而且在所有线性函数中，最小二乘估计的方差最小. (高斯-马尔科夫定理)

结论：

- $\hat{y}_0$ 是 $E(y_0) = \beta_0 + \beta_1 x_0$ 的无偏估计，不表示y_0的估计，因为y_0是随机变量，它不能被估计，但对其可以作预测。
- 除 $\bar{x} = 0$ 外, $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 是相关的.
- 要提高 $\hat{\beta}_0, \hat{\beta}_1$ 的估计精度 (即降低它们的方差) 就要求 n 大, l_{xx} 大 (即要求 $x_1, x_2, \cdots, x_n$ 较分散).

证明

(1) 可把 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 改写为

$$\hat{\beta}_1 = \frac{l_{xy}}{l_{xx}} = \sum \frac{x_i - \bar{x}}{l_{xx}} y_i,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}}\right] y_i.$$

$\hat{\beta}_1$ 和 $\hat{\beta}_0$是独立正态变量 $y_1, y_2, \cdots, y_n$ 的线性组合, 故都服从正态分布,

其期望与方差: (利用 $\sum(x_i - \bar{x}) = 0$)

$$E(\hat{\beta}_1) = \sum \frac{x_i - \bar{x}}{l_{xx}} E(y_i) = \sum \frac{x_i - \bar{x}}{l_{xx}} (\beta_0 + \beta_1 x_i) = \beta_1,$$

$$\mathrm{Var}(\hat{\beta}_1) = \sum \left(\frac{x_i - \bar{x}}{l_{xx}}\right)^2 \mathrm{Var}(y_i) = \sum \frac{(x_i - \bar{x})^2}{l_{xx}^2} \sigma^2 = \frac{\sigma^2}{l_{xx}},$$

$$E(\hat{\beta}_0) = E(\bar{y}) - E(\hat{\beta}_1)\bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0,$$

$$\mathrm{Var}(\hat{\beta}_0) = \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}}\right]^2 \mathrm{Var}(y_i) = \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right)\sigma^2,$$

(2) 考虑到诸 y_i 之间的独立性, 可得

$$\mathrm{Cov}(\hat\beta_0,\hat\beta_1)=\mathrm{Cov}\left(\sum\left[\frac{1}{n}-\frac{(x_i-\bar x)\bar x}{l_{xx}}\right]y_i\ ,\sum\frac{x_i-\bar x}{l_{xx}}y_i\right)$$

$$=\sum\ \left[\frac{1}{n}-\frac{(x_i-\bar x)\bar x}{l_{xx}}\right]\frac{x_i-\bar x}{l_{xx}}\boldsymbol{\sigma}^2=-\frac{\bar x}{l_{xx}}\boldsymbol{\sigma}^2$$

(3) $\hat y_0=\hat\beta_0+\hat\beta_1x_0$ 也是 $y_1,y_2,\cdots,y_n$ 的线性组合, 它也服从正态分布, 其期望与方差:

$$E(\hat y_0)=E(\hat\beta_0)+E(\hat\beta_1)x_0=\beta_0+\beta_1x_0=E(y_0)$$

$$\mathrm{Var}(\hat y_0)=\mathrm{Var}(\hat\beta_0)+\mathrm{Var}(\hat\beta_1)x_0^2+2\mathrm{Cov}(\hat\beta_0,\hat\beta_1)x_0$$

$$=\left[\left(\frac{1}{n}+\frac{\bar x^2}{l_{xx}}\right)+\frac{x_0^2}{l_{xx}}-2\frac{x_0\bar x}{l_{xx}}\right]\sigma^2=\left[\frac{1}{n}+\frac{(x_0-\bar x)^2}{l_{xx}}\right]\sigma^2$$

$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X$是从观察值得到的回归方程，它会随观察结果的不同而改变，只反映了由**X**的变化引起的**Y**的变化，没有包含误差项.

问题:

(1) σ^2的点估计是什么?

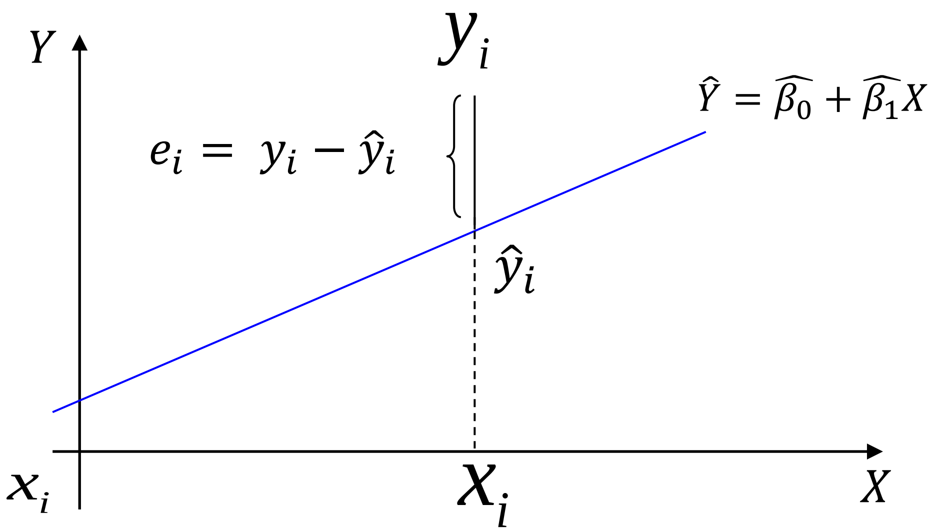
(2) 回归方程是否有意义, 即自变量 X 的变化是否真的对因变量 Y 有影响? 需要对回归效果作出检验.

(3) 如果方程真有意义，用它预测 Y 时，能否估计预测值与真值的偏差?

三、σ^2 的点估计

$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$, $e_i = y_i - \hat{y}_i$称为 $\boldsymbol{x_i}$ 处的残差

$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2$称为残差平方和



$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 反映了除 X 外其它因素对 Y的影响, 这些其它因素没有反映在自变量X中, 可作为随机因素看待.

可以证明 $\frac{Q_e}{\sigma^2} \sim \chi^2(n-2)$

因此 $E\left(\frac{Q_e}{\sigma^2}\right) = n-2 \quad \Rightarrow \quad E\left(\frac{Q_e}{n-2}\right) = \sigma^2$

$\boxed{\hat{\sigma}^2 = \frac{Q_e}{n-2}}$ 是 σ^2 的无偏估计.

四.线性回归的显著性检验

平方和分解公式 $S_T = S_R + Q_e$

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

离差平方和S_T

回归平方和S_R

残差平方和Q_e

0

$$S_R = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \hat{\beta}_1^2 l_{xx}, \qquad Q_e = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$\frac{S_T}{\sigma^2} \sim \chi^2(n-1)$, $\frac{Q_e}{\sigma^2} \sim \chi^2(n-2)$, 得出 $\frac{S_R}{\sigma^2} \sim \chi^2(1)$

$$F = \frac{\frac{S_R}{\sigma^2}/1}{\frac{Q_e}{\sigma^2}/(n-2)} = \frac{S_R}{Q_e/(n-2)} \sim F(1, n-2)$$

$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)=0$ 的证明

$\hat{\beta}_0,\hat{\beta}_1$ 满足正规方程组, 因此有

$$\sum\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0\Rightarrow\sum(y_i-\hat{y}_i)=0,$$

$$\sum\ \left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)x_i=0\Rightarrow\sum(y_i-\hat{y}_i)x_i=0.$$

利用 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i=\bar{y}+\hat{\beta}_1(x_i-\bar{x})$, 可得

$$\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum(y_i-\hat{y}_i)\left[\hat{\beta}_1(x_i-\bar{x})\right]$$

$$=\hat{\beta}_1\left[\sum(y_i-\hat{y}_i)x_i-\sum(y_i-\hat{y}_i)\bar{x}\right]=0$$

对检验问题 $H_0: \beta_1 = 0 \quad \Leftrightarrow \quad H_1: \beta_1 \neq 0$

取检验统计量$F = \frac{S_R}{Q_e/(n-2)} = \frac{\hat{\beta}_1^2 l_{xx}}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}$,

当 H_0 不成立时，F的值应偏大

设显著性水平α,则拒绝域$W = \{F > F_\alpha(1, n-2)\}$

（1）当 $F > F_\alpha$ 时，拒绝 H_0，即可认为变量 y与 x 有线性相关关系；

（2）当 $F \le F_\alpha$ 时，接受 H_0，即可认为变量 y 与 x 没有线性相关关系；

五. 利用回归方程进行预测

- **点预测**

对给定的 $\boldsymbol{X} = \boldsymbol{x}_0$，利用回归方程 $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X$ 可以作出 Y_0 的点预测值 $\hat{Y}_0 = \widehat{\beta_0} + \widehat{\beta_1}x_0$

- **区间预测**

设 X = x_0 时，Y_0 的置信度为 $1-\alpha$ 的预测区间为

$$\left(\hat{Y}_0 - \delta(x_0)\,,\quad \hat{Y}_0 + \delta(x_0)\right)$$

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

$$\hat{\sigma} = \sqrt{\frac{Q_e}{n-2}}$$

推导

$$\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0 \sim N\left(\beta_0+\beta_1 x_0, \sigma^2 \underbrace{\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]}_{*}\right)$$

$\frac{\hat{Y}_0-E(Y_0)}{\sigma\sqrt{*}} \sim N(0,1)$ 但 σ 未知.

$\hat{\sigma}^2=\frac{Q_e}{n-2}$, $\frac{Q_e}{\sigma^2}=\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$,

$$\frac{\hat{Y}_0-E(Y_0)}{\hat{\sigma}\sqrt{*}}=\frac{\frac{\hat{Y}_0-E(Y_0)}{\sigma\sqrt{*}}}{\frac{\hat{\sigma}}{\sigma}}=\frac{Z}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}}=\frac{Z}{\sqrt{\frac{(n-2)\hat{\sigma}^2/\sigma^2}{n-2}}}=\frac{Z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}} \sim t(n-2)$$

故给定显著性水平α, $E(Y_0)$的置信水平为 $1-\alpha$ 的**置信区间**为 $[\hat{Y}_0-\delta(x_0), \hat{Y}_0+\delta(x_0)]$,

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

$Y_0 = E(Y_0) + \varepsilon \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\right)$

Y_0与$\hat{Y}_0$相互独立

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2 \underbrace{\left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{l_{xx}}\right]}_{*}\right)$$

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma}\sqrt{1 + 1/n + (x_0 - \bar{x})^2/l_{xx}}} \sim t(n-2),$$

故给定显著性水平α, Y_0的置信水平为 $1-\alpha$ 的**预测区间**为$\left[\hat{Y}_0 - \delta(x_0), \hat{Y}_0 + \delta(x_0)\right]$,

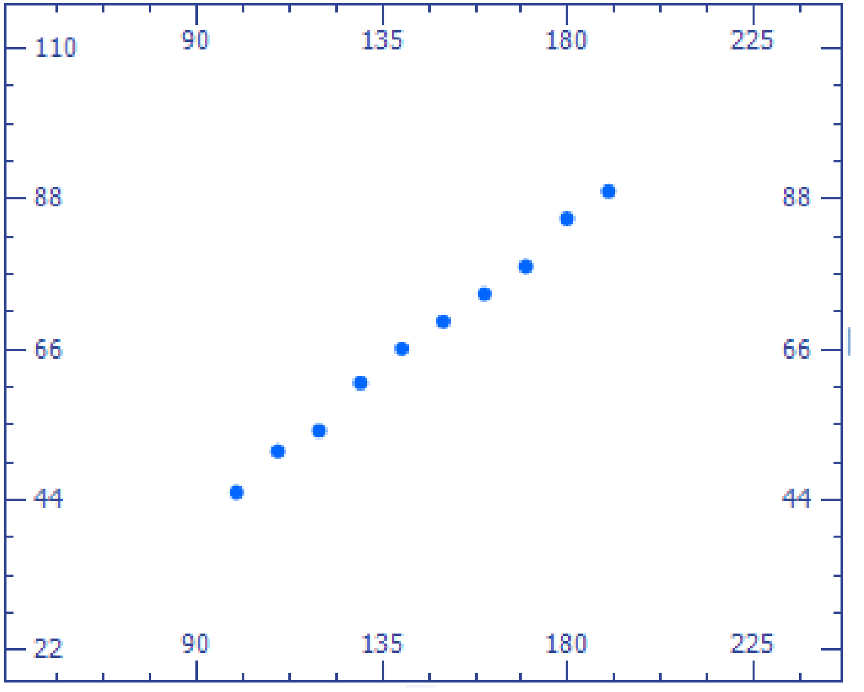
其中 $\delta(x_0) = \hat{\sigma}\sqrt{1 + 1/n + (x_0 - \bar{x})^2/l_{xx}}\, t_{\alpha/2}(n-2), \hat{\sigma} = \sqrt{Q/(n-2)}.$

例 为研究某一化学反应过程温度对产品得率的影响，测得观测数据如下:

温度X/℃	100	110	120	130	140	150	160	170	180	190
得率Y /%	45	51	54	61	66	70	74	78	85	89

求Y关于X的回归方程 $f(X)$，并进行有关检验和预测.

解 先画出散点图.



大致呈线性关系，即有关系 $f(X)=\beta_0+\beta_1X$

计算有关数据 $n=10,\ \sum_{i=1}^{10} x_i y_i = 101570$

$$\sum_{i=1}^{10} x_i = 1450,\ \overline{x} = 145,\ \sum_{i=1}^{10} x_i^2 = 218500$$

$$\sum_{i=1}^{10} y_i = 673,\ \overline{y} = 67.3,\ \sum_{i=1}^{10} y_i^2 = 47225$$

$$l_{xx} = \sum_{i=1}^{10} x_i^2 - n(\overline{x})^2 = 218500 - 10 \times 145^2 = 8250$$

$$l_{xy} = \sum_{i=1}^{10} x_i y_i - n\overline{x}\overline{y} = 101570 - 10 \times 67.3 \times 145 = 3985$$

(1) 回归系数与回归直线方程

回归系数

$$\widehat{\beta_1} = \frac{l_{xy}}{l_{xx}} = \frac{3985}{8250} = 0.48303030$$

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = 67.3 - 0.48303030 \times 145 = -2.73939393$$

所以**回归直线方程**为

$$\hat{Y} = -2.73939393 + 0.48303030 \cdot X$$

或者为 $\hat{Y} = 67.3 + 0.48303030 \cdot (X - 145)$

(2) 线性回归的显著性检验（α=0. 01）

检验问题 $H_0: \beta_1 = 0 \quad \Leftrightarrow \quad H_1: \beta_1 \neq 0$

$$F = \frac{S_R}{Q_e/(n-2)} = \frac{1924.875757}{7.224243/(10-2)}$$

$$= 2131.573591 >> F_{0.01}(1, 10-2) = 11.26$$

故回归效果非常显著

(3) 在 $x_0 = 145$ 进行点预测和区间预测($\alpha = 0.05$)

- **点预测值**

$$\hat{y}_0 = \widehat{\beta_0} + \widehat{\beta_1} x_0 = -2.73939393 + 0.48303030 \times 145 = 67.3$$

- **区间预测值**

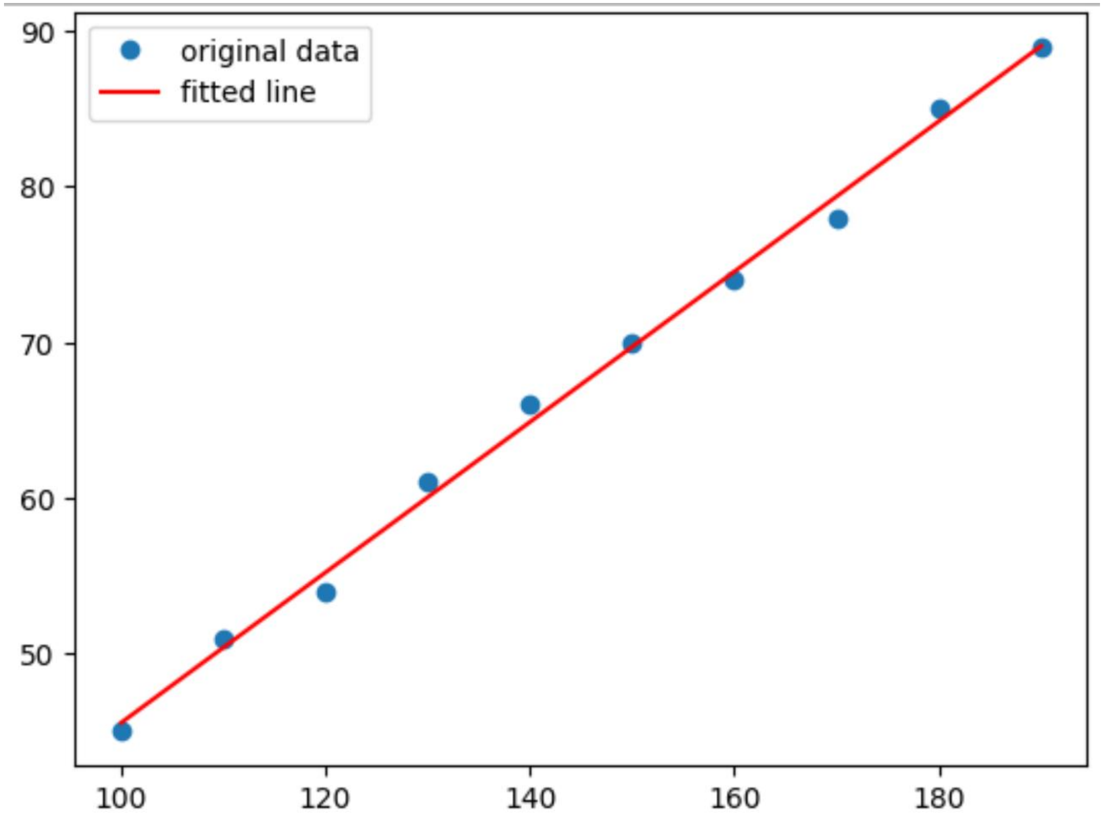
$$\delta(x_0) = \delta(145) = t_{\alpha/2}(n-2) \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{l_{xx}}}$$

$$= 2.306 \times \sqrt{0.903030375} \times \sqrt{1 + \frac{1}{10}} = 2.2983$$

$$(\hat{y}_0 - \delta(x_0),\ \hat{y}_0 + \delta(x_0)) = (65.0,\ 69.60)$$

```
import numpy as np
import matplotlib.pyplot as plt
plt.style.use('default')
from scipy import stats

x = np.array([…])
y = np.array([…])
# Perform the linear regression:
res = stats.linregress(x, y)

plt.plot(x, y, 'o', label='original data')
plt.plot(x, res.intercept + res.slope*x, 'r', label='fitted line')
plt.legend()
plt.show()

print(f"R-squared: {res.rvalue**2:.6f}")

# Calculate 95% confidence interval on slope and intercept:
# Two-sided inverse Students t-distribution
# p - probability, df - degrees of freedom
from scipy.stats import t
tinv = lambda p, df: abs(t.ppf(p/2, df))
ts = tinv(0.05, len(x)-2)
```

```
print(f"slope (95%): {res.slope:.6f} +/- {ts*res.stderr:.6f} \
= [{res.slope - ts*res.stderr:.6f}, {res.slope + ts*res.stderr:.6f}]")
print(f"intercept (95%): {res.intercept:.6f} +/-
{ts*res.intercept_stderr:.6f} \
= [{res.intercept - ts*res.intercept_stderr:.6f}, {res.intercept +
ts*res.intercept_stderr:.6f}]")

y_predictions = res.slope * x + res.intercept
Se = np.sum((y - y_predictions)**2)
Sr = np.sum((y_predictions - y.mean())**2)
n = len(x)

# 假设检验 res.slope = 0
t = np.sqrt(Sr/(Se/(n-2)))
print(f"t (5%)-分位数: {ts:.6f}, 检验统计量 t: {t:.6f}")

#prediction
new_x = …
new_y = res.slope * new_x + res.intercept
x_mean = np.mean(x)
lxx = ((x - x_mean)**2).sum()
delta = np.sqrt(Se/(n-2))* np.sqrt(1 + 1.0 / n + (new_x - x_mean)**2 / lxx)
print(f"new_y (95%): {new_y:.6f} +/- {ts*delta:.6f} \
= [{new_y-ts*delta:.6f}, {new_y+ts*delta:.6f}]")
```



R-squared: 0.996261

slope (95%): 0.483030 +/- 0.024126 = [0.458904, 0.507156]

intercept (95%): -2.739394 +/- 3.566235 = [-6.305629, 0.826841]

t (5%)-分位数: 2.306004, 检验统计量 t: 46.168970

new_y (95%): 67.300000 +/- 2.298305 = [65.001695, 69.598305]

六、参数 $\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\sigma^2}$ 的最大似然估计

一元线性回归模型$Y=\beta_0+\beta_1 X+\varepsilon$, 假定随机项$\varepsilon \sim N\left(0, \sigma^2\right)$.

设 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 是一组相互独立的样本观测值,

由于 $\varepsilon \sim N\left(0, \sigma^2\right), Y=\beta_0+\beta_1 X+\varepsilon \sim N\left(\beta_0+\beta_1 X, \sigma^2\right)$

所以 $Y_i=\beta_0+\beta_1 X_i+\varepsilon_i \sim N\left(\beta_0+\beta_1 X_i, \sigma^2\right)$.

似然函数为 $L\left(\beta_0, \beta_1, \sigma^2\right)=\frac{1}{(\sqrt{2\pi}\sigma)^n} \mathrm{e}^{-\frac{\sum_{i=1}^n (y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}}$,

对数似然函数为

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

根据 $\frac{\partial \ln L\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \sigma^2\right)}{\partial \beta_0}=\frac{\partial \ln L\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \sigma^2\right)}{\partial \beta_1}=\frac{\partial \ln L\left(\beta_0, \beta_1, \sigma^2\right)}{\partial \sigma^2}=0$， 解得

$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$, $\hat{\beta}_1=\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2}$, $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^n \left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2$

填空题 1分

设置

一元线性回归模型中，**σ^2的最大似然估计是** [填空1] **(有/无)偏估计。**

作答

3 多元线性回归分析

一、线性统计模型

设因变量(目标值)Y与自变量(特征值) $x_1, x_2, \cdots, x_k$之间有**线性关系:**

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

这是**线性统计模型**.

假设Y和ε是随机变量，自变量$x_1, x_2, \cdots, x_k$是非随机变量。

作 n 次观测, 得到数据 $(x_{i1}, x_{i2}, \cdots, x_{ik}; y_i), i = 1,2,\cdots,n$, 其中 $y_1, y_2, \cdots, y_n$ 分别是 Y 的 n 次观测值. 若记 $(Y_1, Y_2, \cdots, Y_n)$ 是取自总体 Y 的一个容量为 n 的样本, 则 $(y_1, y_2, \cdots, y_n)$ 是样本 $(Y_1, Y_2, \cdots, Y_n)$ 的观测值,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i,\ i = 1,2,\cdots,n.$$

用向量和矩阵形式表示为

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix},$$

可表示为 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,

假定随机误差 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 是无偏、等方差和不相关的，服从 n 维正态分布：

$$E(\boldsymbol{\varepsilon}) = \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n,$$

其中 σ^2 是末知参数，$\boldsymbol{I}_n$ 是 n 阶单位矩阵，

$E(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2$;

$\mathrm{Cov}\left(\varepsilon_i, \varepsilon_j\right) = 0(i \neq j; i, j = 1,2, \cdots, n)$.

正态线性模型$\left(\boldsymbol{Y}, \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n\right)$

$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

$E(\boldsymbol{Y}) = \boldsymbol{X\beta}$

$\mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{Y}) = \sigma^2 \boldsymbol{I}_n$

统计推断问题:

- 对末知参数向量 $\boldsymbol{\beta}$ 和末知参数 σ^2 进行估计;
- 对 $\boldsymbol{Y}$ 服从线性模型的假设和有关 $\boldsymbol{\beta}$ 的某些假设进行检验;
- 对 $\boldsymbol{Y}$ 进行预测.

假定 $n > k$, 且 $k+1$ 阶方阵 $\boldsymbol{L} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 是可逆矩阵.

二、最小二乘估计

误差平方和 $Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$

最小二乘估计值 $\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} Q(\boldsymbol{\beta})$

正规方程组$\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$，即 $\boldsymbol{L}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$，记 $\boldsymbol{L} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$

$\widehat{\boldsymbol{\beta}} = \boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$,

$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

残差向量$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$

最小二乘估计的性质

性质1 $\widehat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的线性无偏估计量, 且 $\mathrm{Cov}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{L}^{-1}$.

证 由于 $\widehat{\boldsymbol{\beta}} = \boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, 所以 $\hat{\beta}_j(j = 0,1,\cdots,k)$ 是样本 $(Y_1,Y_2,\cdots,Y_n)$ 的线性函数, 这种估计称为**线性估计**, $\widehat{\boldsymbol{\beta}} = \left[\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k\right]^{\mathrm{T}}$ 是 $\boldsymbol{\beta}$ 的线性估计量.

因 $E(\widehat{\boldsymbol{\beta}}) = E\left(\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\right) = \boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}E(\boldsymbol{Y}) = \boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{\beta}$, 故 $\widehat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的线性无偏估计量.

记 $\boldsymbol{G} = \boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}$, 则 $\widehat{\boldsymbol{\beta}} = \boldsymbol{G}\boldsymbol{Y}$. 于是

$$
\begin{aligned}
\mathrm{Cov}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}) &= \mathrm{Cov}(\boldsymbol{G}\boldsymbol{Y},\boldsymbol{G}\boldsymbol{Y}) = \boldsymbol{G}\cdot\mathrm{Cov}(\boldsymbol{Y},\boldsymbol{Y})\cdot\boldsymbol{G}^{\mathrm{T}} = \sigma^2\boldsymbol{G}\boldsymbol{I}_n\boldsymbol{G}^{\mathrm{T}} = \sigma^2\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}} \\
&= \sigma^2\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}} = \sigma^2\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{L}^{-1})^{\mathrm{T}} = \sigma^2\boldsymbol{L}^{-1}.
\end{aligned}
$$

性质2 对于残差向量 $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$, 有

(1) $E(\boldsymbol{e}) = \boldsymbol{0}$;

(2) $\mathrm{Cov}(\boldsymbol{e}, \boldsymbol{e}) = \sigma^2\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\right)$;

(3) $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \boldsymbol{e}) = \boldsymbol{O}$.

性质3 记 $Q_e = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$, 则有 $E(Q_e) = (n-k-1)\sigma^2$,

即 $E\left(\frac{Q_e}{n-k-1}\right) = \sigma^2$, $\hat{\sigma}_e^2 = \frac{Q_e}{n-k-1}$ 是σ^2的无偏估计。

证　由于 $E(\boldsymbol{e})=\boldsymbol{0}$,

$$Q_e=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}=(\boldsymbol{e}-E(\boldsymbol{e}))^{\mathrm{T}}(\boldsymbol{e}-E(\boldsymbol{e}))=\operatorname{tr}\left[(\boldsymbol{e}-E(\boldsymbol{e}))(\boldsymbol{e}-E(\boldsymbol{e}))^{\mathrm{T}}\right].$$

$$E(Q_e)=E\left\{\operatorname{tr}\left[(\boldsymbol{e}-E(\boldsymbol{e}))(\boldsymbol{e}-E(\boldsymbol{e}))^{\mathrm{T}}\right]\right\}$$

$$=\operatorname{tr}\left[E\left([\boldsymbol{e}-E(\boldsymbol{e})][\boldsymbol{e}-E(\boldsymbol{e})]^{\mathrm{T}}\right)\right]$$

$$=\operatorname{tr}[\operatorname{Cov}(\boldsymbol{e},\boldsymbol{e})]=\operatorname{tr}\left[I_n-\boldsymbol{X}\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\right]\sigma^2$$

$$=\left[\operatorname{tr}I_n-\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\right)\right]\sigma^2$$

$$=\left[\operatorname{tr}\boldsymbol{I}_n-\operatorname{tr}\left(\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\right]\sigma^2$$

$$=(\operatorname{tr}\boldsymbol{I}_n-\operatorname{tr}\boldsymbol{I}_{k+1})\,\sigma^2$$

$$=(n-k-1)\sigma^2$$

记$\hat{\sigma}_e^2=\frac{Q_e}{n-k-1}$, $\hat{\sigma}_e^2$ 是未知参数 σ^2 的无偏估计量.

性质4 设 $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}$ 是待估函数, 其中 $\boldsymbol{c}=[c_0,c_1,\cdots,c_k]^{\mathrm{T}}$ 是任一已知的常数向量, 则 $\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ 是 $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}$ 的最小方差线性无偏估计量. 在这个意义下, 称 $\widehat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的**最小方差线性无偏估计**.

性质5 若$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$是正态线性模型, 则

- $\widehat{\boldsymbol{\beta}}$ 与 $\boldsymbol{e}$ 相互独立, 从而 $\widehat{\boldsymbol{\beta}}$ 与 Q_e 相互独立;
- $\widehat{\boldsymbol{\beta}}$ 服从 $k+1$ 维正态分布, $\widehat{\boldsymbol{\beta}}\sim N\left(\boldsymbol{\beta};\sigma^2\boldsymbol{L}^{-1}\right)$;
- $\boldsymbol{e}$ 服从 n 维正态分布, $\boldsymbol{e}\sim N\left(\boldsymbol{0};\sigma^2\left(\boldsymbol{I}_n-\boldsymbol{X}\boldsymbol{L}^{-1}\boldsymbol{X}^{\mathrm{T}}\right)\right)$;
- $\frac{Q_e}{\sigma^2}$ 服从自由度为 $n-k-1$ 的 χ^2 分布, 即

$$\frac{Q_e}{\sigma^2}\sim\chi^2(n-k-1).$$

性质6 若 $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 是正态线性模型, 则 $\widehat{\boldsymbol{\beta}}$ 和 $\hat{\sigma}^2=\frac{Q_e}{n}$ 分别是 $\boldsymbol{\beta}$ 和 σ^2 的极大似然估计量.

三、拟合程度的评价指标

总离差平方和（Total Sum of Squares）

$$SS=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

回归平方和（Explained Sum of Squares）

$$SS_r=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$$

残差平方和（Residual Sum of Squares）

$$SS_e=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=Q_e$$

(平方和分解公式)

在线性模型 $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ 中, $SS=SS_r+SS_e$

R-square(决定系数)

$$R^2 = \frac{SS_r}{SS} = 1 - \frac{SS_e}{SS}$$

是样本回归直线与样本观测值之间的拟合程度的判定指标。

R^2介于0~1之间，越接近1，回归拟合效果越好，一般认为超过0.8的模型拟合优度比较高。

Adjusted R-Square (校正决定系数)

数据集的特征数量越大，R^2 越大，不同数据集的模型结果比较会有一定的误差，引入校正决定系数

$$R_{adj}^2 = 1 - \frac{(n-1)\left(1-R^2\right)}{n-k-1}$$

其中 n 为样本数量，k为特征数量，修正R^2相当于给变量的个数加惩罚项。

四、回归模型的总体显著性假设检验

检验假设 $\quad \mathrm{H}_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0.$

由于 $Y_i \sim N\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2\right), i = 1,2,\cdots,n,$ 相互独立, 所以当 H_0 成立时, $(Y_1, Y_2, \cdots, Y_n)$ 可以看成是取自正态总体 $Y \sim N\left(\beta_0; \sigma^2\right)$ 的一个容量为 n 的样本, $\bar{Y}$ 是样本均值, 从而 $\frac{1}{\sigma^2} SS \sim \chi^2(n-1).$

又 $\frac{1}{\sigma^2} SS_e \sim \chi^2(n-k-1).$

可以证明, SS_r 与 SS_e 相互独立, 因此由 χ^2 分布的可加性,

在 H_0 成立时, $\frac{1}{\sigma^2} SS_r \sim \chi^2(k).$

取检验统计量

$$F = \frac{SS_r/k}{SS_e/(n-k-1)},$$

当 H_0 成立时, $F \sim F(k, n-k-1)$,

当H_0 不成立时, 由于

$$\mathrm{SS}_r = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left[\sum_{j=1}^{k} \hat{\beta}_j \left(x_{ij} - \bar{x}_j\right)\right]^2,$$

F 值有变大的趋势, 因此由

$$F = \frac{SS_r/k}{SS_e/(n-k-1)} > F_\alpha(k, n-k-1)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

方差分析表

方差来源	平方和	自由度	均方和	F 值
回归	$SS_r = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$	k	$MS_r = \frac{SS_r}{k}$	$F = \frac{M_M}{MS_e}$
残差	$SS_e = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	$n - k - 1$	$MS_e = \frac{SS_e}{n-k-1}$	
总和	$SS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n - 1$		

五. 利用回归方程进行预测

对于正态线性模型 $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, 求得 $\boldsymbol{\beta}$ 的最小二乘估计值 $\widehat{\boldsymbol{\beta}}$ 后, 可以用

$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_kx_k$对$Y$进行预测.

记 $\boldsymbol{X}_t=[1,x_{t1},\cdots,x_{tk}]^{\mathrm{T}}$, $\hat{Y}_t=\boldsymbol{X}_t^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$

$\hat{Y}_t\sim N\left(\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta};\sigma^2\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{L}^{-1}\boldsymbol{X}_t\right)$, $Y_t\sim N\left(\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta};\sigma^2\right)$, 所以

$$Y_t-\hat{Y}_t\sim N\left(0;\sigma^2\left(1+\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{L}^{-1}\boldsymbol{X}_t\right)\right).$$

由于 Y_t 与 $\boldsymbol{Y}$ 相互独立, 所以 Y_t 与 Q_e 相互独立.

$\widehat{\boldsymbol{\beta}}$ 与 Q_e 相互独立, 故 $\hat{Y}_t=\boldsymbol{X}_t^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ 与 Q_e 相互独立.

于是, $Y_t-\hat{Y}_t$ 与 Q_e 相互独立.

$$Y_t - \hat{Y}_t \sim N\left(0; \sigma^2\left(1 + \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{L}^{-1} \boldsymbol{X}_t\right)\right)$$

$Q_e/\sigma^2 \sim \chi^2(n-k-1)$,

$Y_t - \hat{Y}_t$ 与 Q_e 相互独立，有

$$\frac{\dfrac{Y_t - \hat{Y}_t}{\sigma\sqrt{1 + \boldsymbol{X}_t^T \boldsymbol{L}^{-1} \boldsymbol{X}_t}}}{\dfrac{1}{\sigma}\sqrt{\dfrac{Q_e}{n-k-1}}} = \frac{Y_t - \hat{Y}_t}{\hat{\sigma}_e\sqrt{1 + \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{L}^{-1} \boldsymbol{X}_t}} \sim t(n-k-1)$$

对于给定的置信水平 $1-\alpha$, Y_t 的预测区间的上、下限为

$$\hat{Y}_t \pm \hat{\sigma}_e\sqrt{1 + \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{L}^{-1} \boldsymbol{X}_t}\, t_{\alpha/2}(n-k-1).$$