

实际应用中,一个自变量同时受多个因变量的影响的情况非常普遍。因此考虑将第 2 章中介绍的一元线性回归拓展到多元的情形。包括多个解释变量的回归模型,就称为多元回归模型。多元线性回归分析是一元情况的简单推广,读者应该注意建立二者之间的联系。

3.1 多元线性回归模型

假设因变量 y 与 m 个解释变量 $x_1, x_2, \dots, x_m$ 具有线性相关关系,取 n 组观察值,则总体线性回归模型为

$$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \dots + w_m x_{im} + u_i, \quad i = 1, 2, \dots, n$$

包含 m 个解释变量的总体回归模型也可以表示为

$$E(y | x_{i1}, x_{i2}, \dots, x_{im}) = w_0 + w_1 x_{i1} + w_2 x_{i2} + \dots + w_m x_{im}, \quad i = 1, 2, \dots, n$$

上式表示在给定 $x_{i1}, x_{i2}, \dots, x_{im}$ 的条件下, y 的条件均值或数学期望。特别地,我们称 w_0 是截距, $w_1, w_2, \dots, w_m$ 是偏回归系数。偏回归系数又称为偏斜率系数。例如, w_1 度量了在其他解释变量 $x_2, x_3, \dots, x_m$ 保持不变的情况下, x_1 每变化 1 个单位时, y 的均值 $E(y | x_{i1}, x_{i2}, \dots, x_{im})$ 的变化。换句话说, w_1 给出了其他解释变量保持不变时, $E(y | x_{i1}, x_{i2}, \dots, x_{im})$ 对 x_1 的斜率。

不难发现,多元线性回归模型是以多个解释变量的固定值为条件的回归分析。

同一元线性回归模型一样,多元线性总体回归模型是无法得到的。所以我们只能用样本观察值进行估计。对应于前面给出的总体回归模型可知多元线性样本回归模型为

$$\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_{i1} + \hat{w}_2 x_{i2} + \dots + \hat{w}_m x_{im}, \quad i = 1, 2, \dots, n$$

和

$$y_i = \hat{w}_0 + \hat{w}_1 x_{i1} + \hat{w}_2 x_{i2} + \dots + \hat{w}_m x_{im} + e_i, \quad i = 1, 2, \dots, n$$

其中, $\hat{y}_i$ 是总体均值 $E(y | x_{i1}, x_{i2}, \dots, x_{im})$ 的估计; $\hat{w}_j$ 是总体偏回归系数 w_j 的估计, $j = 1, 2, \dots, m$; 残差项 e_i 是对随机项 u_i 的估计。

对多元线性总体回归方模型可以用线性方程组的形式表示为

$$\begin{cases} y_1 = w_0 + w_1 x_{11} + w_2 x_{12} + \cdots + w_m x_{1m} + u_1 \\ y_2 = w_0 + w_1 x_{21} + w_2 x_{22} + \cdots + w_m x_{2m} + u_2 \\ \vdots \\ y_n = w_0 + w_1 x_{n1} + w_2 x_{n2} + \cdots + w_m x_{nm} + u_n \end{cases}$$

将上述方程组改写成矩阵的形式:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

或者写成如下形式:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{u}$$

上式就是用矩阵形式表示的多元线性总体回归模型。其中 $\mathbf{y}$ 为 n 阶因变量观察值向量; $\mathbf{X}$ 表示 $n \times m$ 阶解释变量的观察值矩阵; $\mathbf{u}$ 表示 n 阶随机扰动项向量; $\mathbf{w}$ 表示 m 阶总体回归参数向量。

同理,可以得到多元线性样本回归模型的矩阵表示为

$$\mathbf{y} = \mathbf{X}\hat{\mathbf{w}} + \mathbf{e}$$

或者

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}$$

其中 $\hat{\mathbf{y}}$ 表示 n 阶因变量回归拟合值向量; $\hat{\mathbf{w}}$ 表示 m 阶回归参数 $\mathbf{w}$ 的估计值向量; $\mathbf{e}$ 表示 n 阶残差向量。

以上各向量的完整形式如下:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, \quad \hat{\mathbf{w}} = \begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \\ \vdots \\ \hat{w}_m \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

显而易见的是,由于解释变量数量的增多,多元线性回归模型的计算要比一元的情况复杂很多。最后与一元线性回归模型一样,为了对回归模型中的参数进行估计,要求多元线性回归模型在满足线性关系之外还必须遵守以下假定。

1. 零均值假定

干扰项 u_i 均值为零,或对每一个 i ,都有 $E(u_i | x_{i1}, x_{i2}, \cdots, x_{im}) = 0$ 。

2. 同方差假定

干扰项 u_i 的方差保持不变,即 $\text{var}(u_i) = \sigma^2$ 。为了进行假设检验,我们通常认为随机扰动(噪声)符合一个均值为 0、方差为 σ^2 的正态分布,即 $u_i \sim N(0, \sigma^2)$ 。

3. 相互独立性

随机扰动项彼此之间都是相互独立的,即 $\text{cov}(u_i, u_j) = 0$, 其中 $i \neq j$ 。

4. 无多重共线性假定

解释变量之间不存在精确的线性关系,即没有一个解释变量可以被写成模型中其余解释变量的线性组合。

3.2 多元回归模型估计

为了建立完整的多元回归模型,我们需要使用最小二乘法对模型中的偏回归系数进行估计,这个过程所用的许多性质与一元情况下一致。

3.2.1 最小二乘估计量

已知多元线性样本回归模型为

$$y_i = \hat{w}_0 + \hat{w}_1 x_{i1} + \hat{w}_2 x_{i2} + \cdots + \hat{w}_m x_{im} + e_i, \quad i = 1, 2, \dots, n$$

于是离差平方和为

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{w}_0 - \hat{w}_1 x_{i1} - \hat{w}_2 x_{i2} - \cdots - \hat{w}_m x_{im})^2$$

现在求估计的参数 $\hat{w}_0, \hat{w}_1, \dots, \hat{w}_m$, 使得离差平方和取得最小值, 于是根据微积分中极值存在的条件, 要解方程组

$$\begin{cases} \frac{\partial \sum e_i^2}{\partial w_0} = -2 \sum (y_i - \hat{w}_0 - \hat{w}_1 x_{i1} - \cdots - \hat{w}_m x_{im}) = 0 \\ \frac{\partial \sum e_i^2}{\partial w_1} = -2 \sum (y_i - \hat{w}_0 - \hat{w}_1 x_{i1} - \cdots - \hat{w}_m x_{im}) x_{i1} = 0 \\ \vdots \\ \frac{\partial \sum e_i^2}{\partial w_m} = -2 \sum (y_i - \hat{w}_0 - \hat{w}_1 x_{i1} - \cdots - \hat{w}_m x_{im}) x_{im} = 0 \end{cases}$$

其解就是参数 $w_0, w_1, \dots, w_m$ 的最小二乘估计 $\hat{w}_0, \hat{w}_1, \dots, \hat{w}_m$ 。

将以上方程组改写成

$$\begin{cases} n\hat{w}_0 + \sum \hat{w}_1 x_{i1} + \sum \hat{w}_2 x_{i2} + \cdots + \sum \hat{w}_m x_{im} = \sum y_i \\ \sum \hat{w}_0 x_{i1} + \sum \hat{w}_1 x_{i1}^2 + \sum \hat{w}_2 x_{i1} x_{i2} + \cdots + \sum \hat{w}_m x_{i1} x_{im} = \sum x_{i1} y_i \\ \vdots \\ \sum \hat{w}_0 x_{im} + \sum \hat{w}_1 x_{im} x_{i1} + \sum \hat{w}_2 x_{im} x_{i2} + \cdots + \sum \hat{w}_m x_{im}^2 = \sum x_{im} y_i \end{cases}$$

这个方程组称为正规方程组。为了把正规方程组改写成矩阵形式, 记系数矩阵为 $\mathbf{A}$, 常数项向量为 $\mathbf{b}$, $\mathbf{w}$ 的估计值向量为 $\hat{\mathbf{w}}$, 即

$$\begin{aligned}
 \mathbf{A} &= \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \cdots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{im} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{im} & \sum x_{im}x_{i1} & \sum x_{im}x_{i2} & \cdots & \sum x_{im}^2 \end{bmatrix} \\
 &= \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2m} & x_{3m} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} = \mathbf{X}^T \mathbf{X} \\
 \mathbf{b} &= \begin{bmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{im}y_i \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \mathbf{X}^T \mathbf{y}
 \end{aligned}$$

其中 $\hat{\mathbf{w}} = (\hat{w}_0, \hat{w}_1, \dots, \hat{w}_m)^T$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ 。所以正规方程组可以表示为

$$\mathbf{A}\hat{\mathbf{w}} = \mathbf{b} \quad \text{或} \quad (\mathbf{X}^T \mathbf{X})\hat{\mathbf{w}} = \mathbf{X}^T \mathbf{y}$$

当系数矩阵可逆时,正规方程组的解为

$$\hat{\mathbf{w}} = \mathbf{A}^{-1} \mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

进而还可以得到

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

令 $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, 则有 $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, $\mathbf{H}$ 是一个 n 阶对称矩阵, 通常称为帽子矩阵。该矩阵的对角线元素记为 h_{ii} , 它给出了第 i 个观测值离其余 $n-1$ 个观测值的距离有多远, 我们通常称其为杠杆率。

3.2.2 多元回归的实例

现在将通过一个实例来演示在 Python 中建立多元线性回归模型的方法。根据经验, 沉淀物吸收能力是土壤的一项重要特征, 因为它会影响杀虫剂和其他各种农药的有效性。在一项实验中, 我们测定了若干组土壤样本的情况, 数据如表 3-1 所示。其中, y 表示磷酸盐吸收指标; x_1 和 x_2 分别表示可提取的铁含量与可提取的铝含量。请根据这些数据建立 y 关于 x_1 和 x_2 的多元线性回归方程。

表 3-1 土壤沉淀物吸收能力采样数据

x_1	61	175	111	124	130	173	169	169	160	244	257	333	199
x_2	13	21	24	23	64	38	33	61	39	71	112	88	54
y	4	18	14	18	26	26	21	30	28	36	65	62	40

在进行一元线性回归分析之前往往会使用散点图来考查一下解释变量与被解释变量之间的线性关系。在进行多元线性回归分析时,也可以采用类似的图形来观察模型中解释变量与被解释变量间的关系,但所采用的统计图形要更复杂一些,它被称为是散点图阵列,如图 3-1 所示。

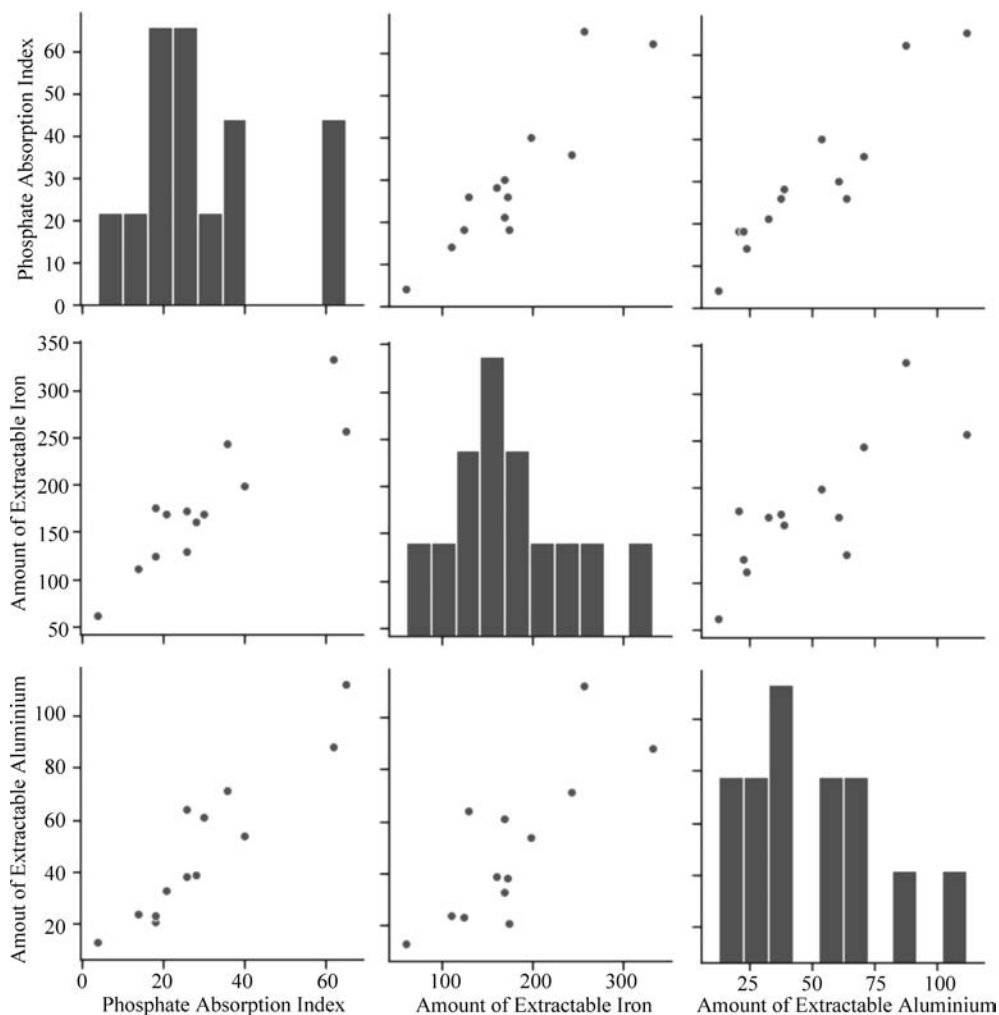


图 3-1 散点图阵列

从散点图中可以看出,每个解释变量都与被解释变量有一定的线性关系,而且这也是我们希望看到的。更重要的是,两个解释变量之间的线性关系并不显著,这就意味着出现多重共线性的可能性较低。在构建多元线性回归模型时,随着解释变量数目的增多,其中某两个解释变量之间产生多重共线性是很容易发生的情况。此时就需要考虑是否将其中某个变量从模型中剔除出去,甚至是重新考虑模型的构建。现在用以检验多重共线性的方法有很多,

有兴趣的读者可以参阅其他相关著作,此处不再赘述。但本书后面还会向读者展示,现代回归分析是如何化解多重共线性的影响的。

下面的代码演示了在 Python 中构建多元线性回归模型的方法,其中用到了 statsmodels 这个库,它是在处理统计学问题时主要用到的一个模块。特别地,要开展基于普通最小二乘法的回归分析,需对 OLS 类进行实例化,而模型的训练过程则通过调用它的 fit() 方法来完成。

```
import numpy as np
import statsmodels.api as sm

pai = np.array([4, 18, 14, 18, 26, 26, 21, 30, 28, 36, 65, 62, 40])
iron = np.array([61, 175, 111, 124, 130, 173, 169, 169, 160, 244, 257, 333, 199])
aluminium = np.array([13, 21, 24, 23, 64, 38, 33, 61, 39, 71, 112, 88, 54])

X_mul = np.vstack((iron, aluminium))
X_mul = X_mul.transpose()
X_mul = sm.add_constant(X_mul)

model_multilin = sm.OLS(pai, X_mul).fit()
model_multilin.summary()
```

执行上述代码,系统将为刚刚构建好的模型输出一份非常全面的报告,如图 3-2 所示。根据图 3-2 所示的结果(虚线框标出的部分),可以写出多元线性回归方程如下:

$$\hat{y} = -7.3507 + 0.1127x_1 + 0.349x_2$$

同时,易见回归模型的拟合优度 $R^2 = 0.948$,调整判定系数 $R_{adj}^2 = 0.938$,说明模型的拟合效果较好。这两个指标的意义在一元的情况下已经进行过详细的介绍,此处不再赘述。需要说明的是,在多元回归的情况下,随着自变量个数的增多,拟合优度也会提高,所以仅仅看这一个指标说服力是有限的。具体这些判定指标的意义,本章后面还会做进一步的解读。

Dep. Variable:	y	R-squared:	0.948
Model:	OLS	Adj. R-squared:	0.938
Method:	Least Squares	F-statistic:	92.03
Date:	Tue, 28 Apr 2020	Prob (F-statistic):	3.63e-07
Time:	20:46:49	Log-Likelihood:	-35.941
No. Observations:	13	AIC:	77.88
Df Residuals:	10	BIC:	79.58
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-7.3507	3.485	-2.109	0.061	-15.115	0.414
x1	0.1127	0.030	3.797	0.004	0.047	0.179
x2	0.3490	0.071	4.894	0.001	0.190	0.508

Omnibus:	1.797	Durbin-Watson:	2.634
Prob(Omnibus):	0.407	Jarque-Bera (JB):	0.679
Skew:	-0.559	Prob(JB):	0.712
Kurtosis:	3.066	Cond. No.	566.

图 3-2 多元线性回归输出结果

3.2.3 总体参数估计量

由 w 的估计量 $\hat{w}$ 的表达式可见, $\hat{w}$ 的每一个分量都是相互独立且服从正态分布的随机变量 $y_1, y_2, \dots, y_n$ 的线性组合, 从而可知随机变量 $\hat{w}$ 服从 $m+1$ 维正态分布。为了求出 $\hat{w}$ 的分布, 首先来计算 $\hat{w}$ 的期望和方差(或方差阵)。

向量 $\hat{w}$ 的数学期望定义为

$$E(\hat{w}) = [E(\hat{w}_0), E(\hat{w}_1), \dots, E(\hat{w}_m)]^T$$

而且对任意 $n \times (m+1)$ 阶矩阵 A , 容易证明

$$E(A\hat{w}) = AE(\hat{w})$$

于是可得

$$\begin{aligned} E(\hat{w}) &= E[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T E(y) \\ &= (X^T X)^{-1} X^T E(Xw + u) = (X^T X)^{-1} X^T Xw = w \end{aligned}$$

所以 $\hat{w}$ 是 w 的无偏估计, 即 $\hat{w}_0, \hat{w}_1, \dots, \hat{w}_m$ 依次是 $w_0, w_1, \dots, w_m$ 的无偏估计, 为了计算 $\hat{w}$ 的方差阵, 先把方差阵写成向量乘积的形式:

$$\begin{aligned} D(\hat{w}) &= \begin{bmatrix} D(\hat{w}_0) & \text{cov}(\hat{w}_0, \hat{w}_1) & \cdots & \text{cov}(\hat{w}_0, \hat{w}_m) \\ \text{cov}(\hat{w}_1, \hat{w}_0) & D(\hat{w}_1) & \cdots & \text{cov}(\hat{w}_1, \hat{w}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{w}_m, \hat{w}_0) & \text{cov}(\hat{w}_m, \hat{w}_1) & \cdots & D(\hat{w}_m) \end{bmatrix} \\ &= E\{[(\hat{w}_0 - E[\hat{w}_0]), (\hat{w}_1 - E[\hat{w}_1]), \dots, (\hat{w}_m - E[\hat{w}_m])]^T \times \\ &\quad [(\hat{w}_0 - E[\hat{w}_0]), (\hat{w}_1 - E[\hat{w}_1]), \dots, (\hat{w}_m - E[\hat{w}_m])]\} \\ &= E\{[\hat{w} - E(\hat{w})][\hat{w} - E(\hat{w})]^T\} \end{aligned}$$

而且^①

$$\begin{aligned} &E\{[\hat{w} - E(\hat{w})][\hat{w} - E(\hat{w})]^T\} \\ &= E\{[(X^T X)^{-1} X^T (y - E(y))][(X^T X)^{-1} X^T (y - E(y))]^T\} \\ &= E[(X^T X)^{-1} X^T (y - E(y))(y - E(y))^T X (X^T X)^{-1}] \\ &= (X^T X)^{-1} X^T E[(y - E(y))(y - E(y))^T] X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T E(uu^T) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

根据已经得到的计算结果, 易知 $\hat{w}$ 的方差阵等于 $\sigma^2 A^{-1}$, 这个方差阵给出了 $\hat{w}$ 中每个元素(即 $\hat{w}_0, \hat{w}_1, \dots, \hat{w}_m$)的方差(或标准差), 以及元素之间的协方差。当 $i=j$ 时, 矩阵对角线上的元素就是相应 $\hat{w}_i$ 的方差 $\text{var}(\hat{w}_i) = \sigma^2 A_{ii}^{-1}$, 由此也可知道 $\hat{w}_i$ 的标准差为

① 计算过程中用到的一些矩阵计算性质如下, 其中 A, B 是两个可以做乘积的矩阵, I 是单位矩阵, 则有 $(AB)^T = B^T A^T, AA^{-1} = I, A^T(A^{-1})^T = (A^{-1}A)^T = I \Rightarrow (A^T)^{-1} = (A^{-1})^T$ 。

$$\text{se}(\hat{w}_i) = \sigma \sqrt{A_{ij}^{-1}}$$

当 $i \neq j$ 时, 矩阵对角线以外的元素就表示相应 $\hat{w}_i$ 与 $\hat{w}_j$ 的协方差, 即 $\text{cov}(\hat{w}_i, \hat{w}_j) = \sigma^2 A_{ij}^{-1}$ 。

例如, 在土壤沉淀物吸收情况的例子中可以求得矩阵 $\mathbf{A}$ 如下:

$$\mathbf{A} = \mathbf{X}^T \mathbf{X} = \begin{bmatrix} 10 & 2305 & 641 \\ 2305 & 467\,669 & 133\,162 \\ 641 & 133\,162 & 41\,831 \end{bmatrix}$$

相应的逆矩阵 $\mathbf{A}^{-1}$ 如下:

$$\mathbf{A}^{-1} = \begin{bmatrix} 0.633\,138 & -0.003\,826 & 0.002\,477 \\ -0.003\,826 & 0.000\,046 & -0.000\,088 \\ 0.002\,477 & -0.000\,088 & 0.000\,265 \end{bmatrix}$$

而且从系统的输出中也知道残差标准误差为 4.379, 于是有

$$\text{se}(\hat{w}_0) = 4.379 \times \sqrt{0.633\,138} \approx 3.484\,37$$

$$\text{se}(\hat{w}_1) = 4.379 \times \sqrt{0.000\,046} \approx 0.029\,69$$

$$\text{se}(\hat{w}_2) = 4.379 \times \sqrt{0.000\,265} \approx 0.071\,30$$

在考虑到计算过程中保留精度存在差异的条件下, 上述参数的标准误差与 3.2.2 节中系统的输出结果是基本一致的。

注意 Python 中的残差标准误差 (Residual Standard Error) 其实就是残差的标准差 (Residual Standard Deviation), 如果读者对于它的计算仍感到困惑, 那么可以参看第 2 章中的相关结论。总的来说, 在多元线性回归中, 总体方差 (同时也是误差项的方差) σ^2 的无偏估计量为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} = \frac{\sum (y_i - \hat{y}_i)^2}{n-k}$$

所以残差 (或误差项) 的标准差为

$$s = \hat{\sigma} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k}}$$

其中, k 是被估计参数的数量。

3.3 从线代角度理解最小二乘

本节将从矩阵的角度来解释最小二乘问题。矩阵方法与微积分方法所得的结论都是一样的, 但是基于矩阵的方法更加强大, 而且更容易推广。在这个过程中, 最关键的内容是理解最佳逼近原理。它是从矩阵 (或线性代数) 角度解释最小二乘问题的核心。

3.3.1 最小二乘问题的通解

非空集合 $S \subset V$, 其中 V 是一个内积空间。定义 $S^\perp = \{x \in V: \langle x, y \rangle = 0, \forall y \in S\}$, 称 $S^\perp$ 为 S 的正交补。

关于这个定义, 有以下两点值得注意:

- (1) 对于 V 的任何子集 S , $S^\perp$ 是 V 的一个子空间。
- (2) 令 S 是 V 的一个子空间, 那么 $S \cap S^\perp = \{0\}$ 。如果 $x \in S \cap S^\perp$, 那么 $\langle x, x \rangle = 0$ 。因此, $x = 0$ 。当然, $0 \in S \cap S^\perp$ 。

注意: 如果 S 不是 V 的一个子空间, 那么第(2)条则不成立。

定理: 令 W 是内积空间 V 的一个有限维子空间, 并令 $y \in V$ 。则有

- (1) $\exists ! u \in W$, 以及 $z \in W^\perp$, 使得 $y = u + z$ 。

- (2) 如果 $\{v_1, v_2, \dots, v_k\}$ 是 W 的一个正交标准基, 那么 $u = \sum_{i=1}^k \langle y, v_i \rangle v_i$ 。

上述定理的几何意义也是非常明确的, 如图 3-3 所示。向量 u 是属于 W 空间的, z 是属于 W 的正交补空间的, 而且 W 和其正交补空间都是子空间。一定存在 u 和 z 使得 $y = u + z$ 。

该定理有诸多非常重要的应用, 其中之一就是给出了下面这个“最佳逼近原理”, 在后面讨论最小二乘问题时, 还会再用到它。

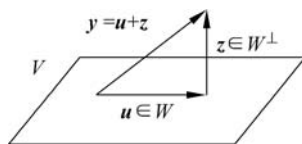


图 3-3 定理的几何意义

假设 W 是 $\mathbb{R}^n$ 空间中的一个子空间, y 是 $\mathbb{R}^n$ 中的任意向量, $\hat{y}$ 是 y 在 W 上的正交投影, 那么 $\hat{y}$ 是 W 中最接近 y 的点, 也就是说, $\|y - \hat{y}\| < \|y - v\|$ 对所有属于 W 又异于 $\hat{y}$ 的 v 成立。这个结论也称为最佳逼近原理。其中, 向量 $\hat{y}$ 称为 W 中元素对 y 的最佳逼近。对于给定元素 y , 可以被某个给定子空间中的元素代替或“逼近”, 我们用 $\|y - v\|$ 表示从 y 到 v 的距离, 则可以认为是用 v 代替 y 的“误差”, 而最佳逼近原理说明误差在 $v = \hat{y}$ 处取得最小值。

最佳逼近原理乍看起来有些抽象, 但其几何意义却是非常直观的, 如图 3-4 所示。不仅如此, 它的证明也非常容易, 只要使用勾股定理即可。

$$\|y - v\|^2 = \underbrace{\|y - \hat{y}\|}_{\in W^\perp}^2 + \underbrace{\|\hat{y} - v\|}_{\in W}^2$$

因为 $y - \hat{y} \perp \hat{y} - v$, 所以应用勾股定理可得

$$\|y - \hat{y} + \hat{y} - v\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - v\|^2 \geq \|y - \hat{y}\|^2$$

当且仅当 $\hat{y} = v$ 时取等号。

现在回到一直在讨论的回归问题上。假设在某些时间点 x_i 上, 观察到一些数值 y_i , 于是有一组二维数据点 (x_i, y_i) , 其中 $i = 1, 2, \dots, n$ 。现在的任务是找到一条最能代表这些数据的直线。如图 3-5 所示, 假设备选直线的方程为 $y = bx + c$ 。在 x_i 时刻, 实际观测值为 y_i , 而根据回归直线, 估计值应该是 $bx_i + c$ 。

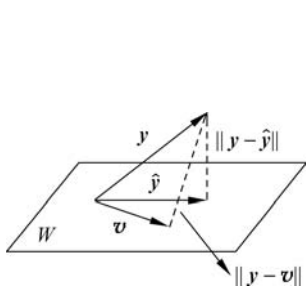


图 3-4 最佳逼近原理的几何解释

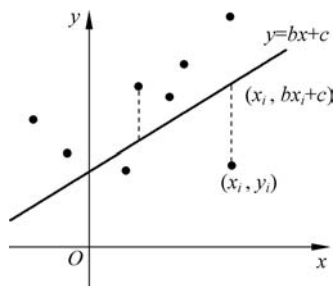


图 3-5 线性回归

因此,实际观测值与估计值之间的误差就可以表示为 $|y_i - cx_i - d|$ 。既然我们的任务是找到那条“最好的”直线,那么就可以用“总误差”最小来作为衡量所谓“最好的”标准。而且等价地,我们用平方来取代绝对值计算,注意 (x_i, y_i) 是已知的,于是现在的目标变为求使得下式(总误差)最小的参数 b 和 c :

$$\sum_{i=1}^n (y_i - bx_i - c)^2$$

一种可以想到的解法就是采用微积分的方法,这也是本书前面一直采用的方法。下面来看看如何把上述优化问题转变为矩阵问题。令

$$\mathbf{X} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} b \\ c \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

此时问题就变成如下形式:

$$E = \sum_{i=1}^n (y_i - cx_i - d)^2 = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

可见这个形式也简单明了。不仅如此,矩阵也非常便于问题的推广,比如说现在不是要找直线,而是一个用多项式所表示的曲线(这时问题就变成了多元回归问题),例如,二次曲线 $y = ax^2 + bx + c$ 。这时其实只要像下面这样简单地改变一下矩阵的形式即可:

$$\mathbf{X} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

事实上,可以看出对于多元回归问题,最终要面对的问题都是要求一个 $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ 的最佳逼近,采用线性代数方法,就将一个相当复杂的问题统一到一个相对简单的形式上来。

假设有一个很大的方程组 $\mathbf{X}\mathbf{w} = \mathbf{y}$, 如果方程组的解不存在,但又需要对它进行求解时,

其实要做的就是去寻找一个最佳的 w , 使得 Xw 尽量接近 y 。

考虑 Xw 作为 y 的一个近似, 那么 y 到 Xw 的距离越小, 以 $\|Xw - y\|$ 来度量的近似程度也就越好。一般的最小二乘问题就是要找出使得 $\|Xw - y\|$ 尽量小的 w , 术语“最小二乘”来源于这样的事实: $\|Xw - y\|$ 是平方和的平方根。

定义: 如果 $m \times n$ 的矩阵 X 和向量 $y \in \mathbb{R}^m$, $Xw - y$ 的最小二乘解是 $\hat{w} \in \mathbb{R}^n$, 使得 $\|X\hat{w} - y\| \leq \|Xw - y\|$ 对所有的 $w \in \mathbb{R}^n$ 成立。

回忆一下列空间的概念。假设大小为 $m \times n$ 的矩阵 $X = [x_1, x_2, \dots, x_n]$, 此处, x_i 表示矩阵中的一列, 其中 $1 \leq i \leq n$, 那么矩阵 X 的列空间(记为 $\text{Col}X$)就是由 X 中各列的所有线性组合构成的集合, 即 $\text{Col}X = \text{Span}\{x_1, x_2, \dots, x_n\}$ 。

很显然, $\text{Col}X = \text{Span}\{x_1, x_2, \dots, x_n\}$ 是一个子空间, 进而可以知道 $m \times n$ 的矩阵 X 的列空间是 $\mathbb{R}^m$ 的一个子空间。此外, 我们还注意到 $\text{Col}X$ 中的一个典型向量(或者称为元素)可以写成 Xw 的形式, 其中 w 是某个向量, 记号 Xw 表示 X 的列向量的一个线性组合, 也就是说, $\text{Col}X = \{y : y = Xw, w \in \mathbb{R}^n\}$, 于是该式告诉我们:

- (1) 记号 Xw 代表 $\text{Col}X$ 空间中的向量;
- (2) $\text{Col}X$ 是线性变换 $w \mapsto Xw$ 的值域。

假设 W 是 $\mathbb{R}^n$ 空间中的一个子空间, y 是 $\mathbb{R}^n$ 中的任意向量, $\hat{y}$ 是 y 在 W 上的正交投影, 那么 $\hat{y}$ 是 W 中最接近 y 的点, 也就是说, $\|y - \hat{y}\| \leq \|y - v\|$ 对所有属于 W 又异于 $\hat{y}$ 的 v 成立。这也就是本节开始时介绍的最佳逼近原理。其中, 向量 $\hat{y}$ 称为 W 中元素对 y 的最佳逼近。还可以知道, 对于给定元素 y , 可以被某个给定子空间中的元素代替或“逼近”, 用 $\|y - v\|$ 表示的从 y 到 v 的距离, 可以认为是用 v 代替 y 的“误差”, 而最佳逼近原理说明误差在 $\hat{y} = v$ 处取得最小值。

最佳逼近原理看似抽象, 其实它的几何意义也是相当直观的。如图 3-6 所示, 所取的向量空间为 $\mathbb{R}^2$, 其中的一个子空间 W 就是横轴, y 是 $\mathbb{R}^2$ 中的任意向量, $\hat{y}$ 是 y 在 W 上的正交投影, 那么 $\hat{y}$ 显然是 W 中最接近 y 的点。更重要的是, 这个结论不仅仅在欧几里得几何空间中成立, 因为泛函中所讨论的“向量”可能是我们通常所说的向量, 但也可能是函数、矩阵, 甚至多项式等, 但最佳逼近原理皆成立。

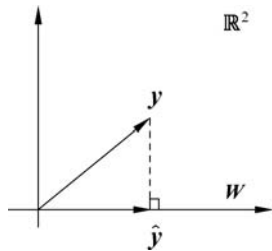


图 3-6 最小二乘法的几何解释

通过前面的分析, 易知, 最小二乘问题的最重要特征是无
论怎么选取 w , 向量 Xw 必然位于列空间 $\text{Col}X$ 中。而最小二乘问题的本质任务就是要寻找一个 w , 使得 Xw 是 $\text{Col}X$ 中最接近 y 的向量。

还可以在复数域中把最小二乘问题重述如下: 给定 $X \in M_{m \times n}(\mathbb{F})$, 以及 $y \in \mathbb{F}^m$, 找到一个 $\hat{w} \in \mathbb{F}^n$, 使得 $\|X\hat{w} - y\|$ 最小(特别地, 对于回归问题而言, 就是给定一个测度的集合 (x_i, y_i) , 其中 $i = 1, 2, \dots, n$, 找到一个最佳的 $n-1$ 阶多项式, 来表示这个测度的集合)。

与之前的情况一样, 只要利用“最佳逼近原理”, 问题即可迎刃而解! 如图 3-7 所示, 子

空间 W 是 X 的值域, 即 W 中的每一个元素都是 Xw 。
 y 是 $\mathbb{F}^m$ 空间中的一个元素, 如果要在 W 中找一个使得
 $\|Xw - y\|$ 最小的 $\hat{w}$, 显然 $\hat{w}$ 就是 y 在 W 中的投影。

特别地, 如果 $\text{rank}(X) = n$, 则有唯一解 $\hat{w} = (X^*X)^{-1}X^*y$ 。这与本书前面讨论的最小二乘解相一致, 但此时我们把结论拓展到一个更大的范围内(复数域)给出结论。

注意到, 令 $w_i = 1, 2, \dots, m$ 都是互不相同的, 假设 $m \geq n$, 则有 $\text{rank}(X) = n$ 。

下面来推导这个结论。我们声称 $\hat{w}$ 是最小二乘问题的解, 意思就是需要证明, 对于任何的 $w \in \mathbb{F}^n$, 有 $\|X\hat{w} - y\| \leq \|Xw - y\|$ 成立。

证明: $\|Xw - y\|^2 = \|X\hat{w} - y + Xw - X\hat{w}\|^2$, 根据勾股定理可得, $\|X\hat{w} - y + Xw - X\hat{w}\|^2 = \|X\hat{w} - y\|^2 + \|Xw - X\hat{w}\|^2 \geq \|X\hat{w} - y\|^2$ 。

对于所有的 $w \in \mathbb{F}^n$,

$$\begin{aligned} y - X\hat{w} \perp W = R(X) &\Rightarrow \langle Xw, y - X\hat{w} \rangle = 0 \\ &\Rightarrow \langle w, X^*(y - X\hat{w}) \rangle = 0 \end{aligned}$$

因为对于所有的 $w \in \mathbb{F}^n$ 上式都成立, 所以有

$$X^*(y - X\hat{w}) = 0 \Rightarrow X^*X\hat{w} = X^*y$$

因为 X^* 的大小是 $n \times m$, X 的大小是 $m \times n$, 所以 $X^*X \in M_{n \times n}(\mathbb{F})$, 如果 $\text{rank}(X^*X) = n$, 那么 X^*X 是可逆的, 则有

$$\hat{w} = (X^*X)^{-1}X^*y$$

上述过程中的几个步骤可以实现, 由下面两个引理来保证。

引理 1: $\langle Xw, y \rangle_m = \langle w, X^*y \rangle_n$, 这里 $X \in M_{m \times n}(\mathbb{F})$, $w \in \mathbb{F}^n$ 以及 $y \in \mathbb{F}^m$ 。

引理 2: $\text{rank}(X^*X) = \text{rank}(X)$, $X \in M_{m \times n}(\mathbb{F})$ 。

综上所述, 结论得证。

3.3.2 最小二乘问题的计算

对于给定的 X 和 y , 运用最佳逼近定理于 $\text{Col}X$ 空间, 取

$$\hat{y} = \text{proj}_{\text{Col}X} y$$

由于 y 属于 X 的列空间, 方程 $Xw = \hat{y}$ 是相容的且存在一个属于 $\mathbb{R}^n$ 的 $\hat{w}$ 使得 $X\hat{w} = \hat{y}$ 。由于 $\hat{y}$ 是 $\text{Col}X$ 中最接近 y 的点, 一个向量 $\hat{w}$ 是 $Xw = y$ 的一个最小二乘解的充分必要条件是 $\hat{w}$ 满足 $X\hat{w} = \hat{y}$ 。这个属于 $\mathbb{R}^n$ 的 $\hat{w}$ 是一系列由 X 构造 $\hat{y}$ 的权值(显然如果方程 $X\hat{w} = \hat{y}$ 中存在自由变量, 则这个方程会有多个解)。

如果 $\hat{w}$ 满足 $X\hat{w} = \hat{y}$, 根据正交分解定理, 投影 $\hat{y}$ 具有这样的性质: $(y - \hat{y})$ 与 $\text{Col}X$ 正交, 即 $(y - X\hat{w})$ 正交于 X 的每一列, 如果 x_j 是 X 的任意列, 那么 $x_j \cdot (y - X\hat{w}) = 0$ 且

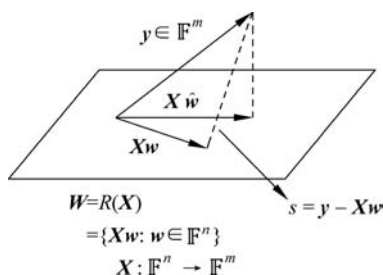


图 3-7 最小二乘的几何解释

$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = 0$ (注意前者是内积运算, 后者是矩阵乘法), 由于每个 $\mathbf{x}_j^T$ 是 $\mathbf{X}^T$ 的一行, 于是可得

$$\mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) = 0$$

因此有

$$\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}^T\mathbf{y} = 0$$

或者

$$\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{y}$$

这其实表明 $\mathbf{X}\mathbf{w} = \mathbf{y}$ 的每个最小二乘解满足方程

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

上述矩阵方程表示的线性方程组常称为 $\mathbf{X}\mathbf{w} = \mathbf{y}$ 的法方程, 其解通常用 $\hat{\mathbf{w}}$ 表示。最后可以得出如下定理(证明从略)。

定理: 方程 $\mathbf{X}\mathbf{w} = \mathbf{y}$ 的最小二乘解集和法方程 $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$ 的非空解集一致。

下面通过一个例子来说明最小二乘问题的解法。

例: 求不相容方程 $\mathbf{X}\mathbf{w} = \mathbf{y}$ 的最小二乘解, 其中

$$\mathbf{X} = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}$$

利用前面给出的公式计算

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \\ \mathbf{X}^T\mathbf{y} &= \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix} \end{aligned}$$

那么法方程 $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$ 就变成了

$$\begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}$$

行变换可用于解此方程组, 但由于 $\mathbf{X}^T\mathbf{X}$ 是 2×2 的可逆矩阵, 于是很容易得到

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{84} \begin{bmatrix} 5 & -1 \\ -1 & 17 \end{bmatrix}$$

那么可以解 $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$ 如下:

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{84} \begin{bmatrix} 5 & -1 \\ -1 & 17 \end{bmatrix} \begin{bmatrix} 19 \\ 11 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

在很多计算中, $\mathbf{X}^T$ 是可逆的, 但也并非都是如此, 下面例子中的矩阵常常出现在统计学中的方差分析问题里。

例: 求 $\mathbf{X}\mathbf{w} = \mathbf{y}$ 的最小二乘解, 其中

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 5 \\ 1 \end{bmatrix}$$

类似地,可以算得

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix}$$

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ -4 \\ 2 \\ 6 \end{bmatrix}$$

矩阵方程 $\mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{X}^T \mathbf{y}$ 的增广矩阵为

$$\begin{bmatrix} 6 & 2 & 2 & 2 & 4 \\ 2 & 2 & 0 & 0 & -4 \\ 2 & 0 & 2 & 0 & 2 \\ 2 & 0 & 0 & 2 & 6 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & -1 & -5 \\ 0 & 0 & 1 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

通解是: $x_1 = 3 - x_4$, $x_2 = -5 + x_4$, $x_3 = -2 + x_4$, x_4 是自由变量。

所以, $\mathbf{X} \mathbf{w} = \mathbf{y}$ 的最小二乘通解具有如下形式:

$$\hat{\mathbf{w}} = \begin{bmatrix} 3 \\ -5 \\ -2 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

在什么条件下,方程 $\mathbf{X} \mathbf{w} = \mathbf{y}$ 的最小二乘解是唯一的? 下面的定理给出了判断准则(当然,正交投影总是唯一的),我们不具体讨论该定理的证明。

定理: 矩阵 $\mathbf{X}^T \mathbf{X}$ 是可逆的,其充分必要条件是 $\mathbf{X}$ 的列是线性无关的,在这种情况下,方程 $\mathbf{X} \mathbf{w} = \mathbf{y}$ 有唯一最小二乘解 $\hat{\mathbf{w}}$ 且它有下面的表达式

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

注意,这是一个通解,也就是多元线性回归问题的解都可以用它来表示。每种具体的线

性回归问题的解都可以看成是它的特例,例如第2章中给出的一元线性回归的解是

$$\hat{w}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\hat{w}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

下面就基于线性回归的通解来推导证明上述关于一元线性回归最小二乘解的结论。对于一元线性回归而言,假设训练集中有 n 个数据:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

代入通解公式 $\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, 则有

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

其中,

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

另外,

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

因此,

$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

综上可得

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\begin{aligned} \begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \end{bmatrix} &= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\ &= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ n \sum x_i y_i - \sum x_i \sum y_i \end{bmatrix} \end{aligned}$$

最终证明

$$\begin{aligned} \hat{w}_0 &= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} \\ \hat{w}_1 &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \end{aligned}$$

下面的例子表明,当 $\mathbf{X}$ 的列向量不正交时,该如何找到 $\mathbf{X}\mathbf{w}=\mathbf{y}$ 的最小二乘解。这类矩阵在线性回归中常被用到。

例:找出 $\mathbf{X}\mathbf{w}=\mathbf{y}$ 的最小二乘解,其中

$$\mathbf{X} = \begin{bmatrix} 1 & -6 \\ 1 & -2 \\ 1 & 1 \\ 1 & 7 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix}$$

解:由于 $\mathbf{X}$ 的列 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 相互正交, $\mathbf{y}$ 在 $\text{Col}\mathbf{X}$ 的正交投影如下:

$$\hat{\mathbf{y}} = \frac{\mathbf{y} \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1} \cdot \mathbf{x}_1 + \frac{\mathbf{y} \cdot \mathbf{x}_2}{\mathbf{x}_2 \cdot \mathbf{x}_2} \cdot \mathbf{x}_2 = \frac{8}{4} \cdot \mathbf{x}_1 + \frac{45}{90} \cdot \mathbf{x}_2 = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} -3 \\ -1 \\ 1/2 \\ 7/2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 5/2 \\ 11/2 \end{bmatrix}$$

既然 $\hat{\mathbf{y}}$ 已知,我们可以解 $\mathbf{X}\hat{\mathbf{w}}=\hat{\mathbf{y}}$ 。这个很容易,因为已经知道 $\hat{\mathbf{y}}$ 用 $\mathbf{X}$ 的列线性(通过线性组合来)表示时的权值。于是从上式可以立刻得到

$$\hat{\mathbf{w}} = \begin{bmatrix} 8/4 \\ 45/90 \end{bmatrix} = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix}$$

某些时候,最小二乘解问题的法方程可能是病态的,也就是 $\mathbf{X}^T\mathbf{X}$ 中元素在计算中出现较小的误差,可导致解 $\hat{\mathbf{w}}$ 出现较大的误差。如果 $\mathbf{X}$ 的列线性无关,最小二乘解常常可通过 $\mathbf{X}$ 的 QR 分解更可靠地求出。来看下面这个定理。

定理: 给定一个 $m \times n$ 的矩阵 $\mathbf{X}$,且具有线性无关的列,取 $\mathbf{X}=\mathbf{Q}\mathbf{R}$ 是 $\mathbf{X}$ 的 QR 分解,那么对每一个属于 $\mathbb{R}^m$ 的 $\mathbf{y}$,方程 $\mathbf{X}\mathbf{w}=\mathbf{y}$ 有唯一的最小二乘解,其解为 $\hat{\mathbf{w}}=\mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}$ 。

这个定理的证明非常简单,这里不再赘述。更进一步,基于 QR 分解的知识,我们知道 $\mathbf{Q}$ 的列形成 $\text{Col}\mathbf{X}$ 的正交基,因此, $\mathbf{Q}\mathbf{Q}^T\mathbf{y}$ 是 $\mathbf{y}$ 在 $\text{Col}\mathbf{X}$ 上的正交投影 $\hat{\mathbf{y}}$,那么 $\mathbf{X}\hat{\mathbf{w}}=\hat{\mathbf{y}}$ 说明 $\hat{\mathbf{w}}$ 是 $\mathbf{X}\mathbf{w}=\mathbf{y}$ 的最小二乘解。

此外,由于上述定理中的 $\mathbf{R}$ 是上三角形矩阵, $\hat{\mathbf{w}}$ 可从方程 $\mathbf{R}\mathbf{w}=\mathbf{Q}^T\mathbf{y}$ 计算得到。求解该方程时,通过回代过程或行变换会比较高效。

例: 求出 $\mathbf{X}\mathbf{w}=\mathbf{y}$ 的最小二乘解, 其中

$$\mathbf{X} = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 0 \\ 1 & 1 & 2 \\ 1 & 3 & 3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 3 \\ 5 \\ 7 \\ -3 \end{bmatrix}$$

解: 可以计算矩阵 $\mathbf{X}$ 的 QR 分解为

$$\mathbf{X} = \mathbf{Q}\mathbf{R} = \begin{bmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & -1/2 \\ 1/2 & -1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 2 & 4 & 5 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix}$$

那么

$$\mathbf{Q}^T\mathbf{y} = \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & -1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 7 \\ -3 \end{bmatrix} = \begin{bmatrix} 6 \\ -6 \\ 4 \end{bmatrix}$$

满足 $\mathbf{R}\mathbf{w}=\mathbf{Q}^T\mathbf{y}$ 的最小二乘解是 $\hat{\mathbf{w}}$, 也就是

$$\begin{bmatrix} 2 & 4 & 5 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -6 \\ 4 \end{bmatrix}$$

这个方程很容易解出得

$$\hat{\mathbf{w}} = \begin{bmatrix} 10 \\ -6 \\ 2 \end{bmatrix}$$

3.4 多元回归模型检验

借由最小二乘法所构建的线性回归模型是否给出了观察值的一种有效描述呢? 或者说, 所构建的模型是否具有一定的解释力? 要回答这些问题, 就需要对模型进行一定的检验。

3.4.1 线性回归的显著性

与一元线性回归类似, 要检测随机变量 y 和可控变量 $x_1, x_2, \dots, x_m$ 之间是否存在线性相关关系, 即检验关系式 $y = w_0 + w_1x_1 + \dots + w_mx_m + u$ 是否成立, 其中 $u \sim N(0, \sigma^2)$ 。此

时主要检验 m 个系数 $w_1, w_2, \dots, w_m$ 是否全为零。如果全为零,则可认为线性回归不显著;若系数 $w_1, w_2, \dots, w_m$ 不全为零,则可认为线性回归是显著的。为进行线性回归的显著性检验,在上述模型中提出原假设和备择假设:

$$H_0: w_1 = w_2 = \dots = w_m = 0$$

$$H_1: H_0 \text{ 是错误的}$$

设对 $(x_1, x_2, \dots, x_m, y)$ 已经进行了 n 次独立观测,得观测值 $(x_{i1}, x_{i2}, \dots, x_{im}, y_i)$, 其中 $i=1, 2, \dots, n$ 。由观测值确定的线性回归方程为

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \dots + \hat{w}_m x_m$$

将 $(x_1, x_2, \dots, x_m)$ 的观测值代入,有

$$\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_{i1} + \dots + \hat{w}_m x_{im}$$

令

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

采用 F 分布检验法。首先对总离差平方和进行分解:

$$SS_{\text{total}} = \sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

与前面在一元线性回归时讨论的一样,残差平方和

$$SS_{\text{residual}} = \sum (y_i - \hat{y}_i)^2$$

反映了试验时随机误差的影响。

回归平方和

$$SS_{\text{regression}} = \sum (\hat{y}_i - \bar{y})^2$$

反映了线性回归引起的误差。

在原假设成立的条件下,可得

$$y_i = w_0 + u_i, \quad i = 1, 2, \dots, n$$

$$\bar{y} = w_0 + \bar{u}$$

观察 $SS_{\text{regression}}$ 和 SS_{residual} 的表达式,易见如果 $SS_{\text{regression}}$ 比 SS_{residual} 大得多,就不能认为所有的 $w_1, w_2, \dots, w_m$ 全为零,即拒绝原假设;反之则接受原假设。从而考虑由这两项之比构造的检验统计量。

由 F 分布的定义可知

$$F = \frac{SS_{\text{regression}}/m}{SS_{\text{residual}}/(n-m-1)} \sim F(m, n-m-1)$$

给定显著水平 α , 由 F 分布表查得临界值 $F_\alpha(m, n-m-1)$, 使得

$$P\{F \geq F_\alpha(m, n-m-1)\} = \alpha$$

由采样得到观测数据,求得 F 统计量的数值,如果 $F \geq F_\alpha(m, n-m-1)$, 则拒绝原假设,即线性回归是显著的;如果 $F < F_\alpha(m, n-m-1)$, 则接受原假设,即认为线性回归方程不显著。

在土壤沉淀物吸收情况的例子中,可以算得 F 统计量的大小为 92.03,这个值要远远大于 $F_{0.05}(2,10)$,所以有理由拒绝原假设,即证明回归总体是显著线性的。也可以通过与这个 F 统计量对应的 P 值来判断,在图 3-2 所示的输出结果中已经给出了该值。或者也可以调用训练所得模型的 `f_pvalue` 属性来输出该值:

```
print(model_multilin.f_pvalue)
```

输出结果为 $3.63\text{e}-07$,这与图 3-2 中给出的结果是一致的。另外,基于对该值计算原理的认识,也可以在 Python 中使用下面的代码得到相应的 P 值:

```
from scipy.stats import f
print(f.sf(92.03, 2, 10))
```

可见, P 值远远小于 0.05,因此有足够的把握拒绝原假设,并同样得到回归总体具有显著线性的结论。

3.4.2 回归系数的显著性

在多元线性回归中,若线性回归显著,回归系数不全为零,则回归方程

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 + \cdots + \hat{w}_m x_m$$

是有意义的。但线性回归显著并不能保证每一个回归系数都足够大,或者说不能保证每一个回归系数都显著地不等于零。若某一系数等于零,如 $w_j = 0$,则变量 x_j 对 y 的取值就不起作用。因此,要考查每一个自变量 x_j 对 y 的取值是否起作用,其中 $j = 1, 2, \cdots, m$,就需对每一个回归系数 w_j 进行检验。为此在线性回归模型上提出原假设:

$$H_0: w_j = 0, \quad 1 \leq j \leq m$$

由于 $\hat{w}_j$ 是 w_j 的无偏估计量,自然由 $\hat{w}_j$ 构造检验用的统计量。由

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

易知 $\hat{w}_j$ 是相互独立的正态随机变量 $y_1, y_2, \cdots, y_n$ 的线性组合,所以 $\hat{w}_j$ 也服从正态分布,并且有

$$E(\hat{w}_j) = w_j, \quad \text{var}(\hat{w}_j) = \sigma^2 A_{jj}^{-1}$$

即 $\hat{w}_j \sim N(w_j, \sigma^2 A_{jj}^{-1})$,其中 A_{jj}^{-1} 是矩阵 $\mathbf{A}^{-1}$ 的主对角线上的第 j 个元素,而且这里的 j 是从第 0 个算起的。于是

$$\frac{\hat{w}_j - w_j}{\sigma \sqrt{A_{jj}^{-1}}} \sim N(0, 1)$$

而

$$\frac{\text{SS}_{\text{residual}}}{\sigma^2} \sim \chi_{n-m-1}^2$$

还可以证明 $\hat{w}_j$ 与 $\text{SS}_{\text{residual}}$ 是相互独立的。因此在原假设成立的条件下,有

$$T = \frac{\hat{w}_j}{\sqrt{A_{jj}^{-1} SS_{\text{residual}} / (n - m - 1)}} \sim t(n - m - 1)$$

给定显著水平 α , 查 t 分布表得到临界值 $t_{\alpha/2}(n - m - 1)$, 由样本值算得 T 统计量的数值, 若 $|T| \geq t_{\alpha/2}(n - m - 1)$ 则拒绝原假设, 即认为 w_j 和零有显著的差异; 若 $|T| < t_{\alpha/2}(n - m - 1)$, 则接受原假设, 即认为 w_j 显著地等于零。

由于 $E[SS_{\text{residual}} / \sigma^2] = n - m - 1$, 所以

$$\hat{\sigma}^{*2} = \frac{SS_{\text{residual}}}{n - m - 1}$$

是 σ^2 的无偏估计。于是 T 统计量的表达式也可以简写为

$$T = \frac{\hat{w}_j}{\hat{\sigma}^* \sqrt{A_{jj}^{-1}}} = \frac{\hat{w}_j}{\text{se}(\hat{w}_j)} \sim t(n - m - 1)$$

在土壤沉淀物吸收情况的例子中, Python 计算得到的各参数估计值为

$$\hat{w}_0 = -7.350\ 66, \quad \hat{w}_1 = 0.112\ 73, \quad \hat{w}_2 = 0.349\ 00$$

于是同 3 个参数相对应的 T 统计量分别为

$$T_{\hat{w}_0} = \frac{-7.350\ 66}{4.379 \times \sqrt{0.633\ 138}} = -2.109$$

$$T_{\hat{w}_1} = \frac{0.112\ 73}{4.379 \times \sqrt{0.000\ 046}} = 3.797$$

$$T_{\hat{w}_2} = \frac{0.349\ 00}{4.379 \times \sqrt{0.000\ 265}} = 4.894$$

相应的 P 值可以在 Python 中用下列代码算得:

```
from scipy.stats import t
print(2 * (t.sf(2.109, 10)))
print(2 * (t.sf(3.797, 10)))
print(2 * (t.sf(4.894, 10)))
```

输出结果如下:

```
0.06114493
0.00350298
0.00062871
```

或者, 也可以直接调用训练所得模型的 `pvalues` 属性来输出该值:

```
print(model_multilin.pvalues)
```

输出结果如下:

```
[0.06110081 0.00350359 0.00062836]
```

以上这些与图 3-2 中输出的结果是一致的(注意, 由于计算精度的问题, 最后几位有效数字

上的差异是正常的),且可据此推断回归系数 w_1 和 w_2 是显著(不为零)的。注意,截距项 w_0 是否为零并不是我们需要关心的。

3.5 多元线性回归模型预测

对于线性回归模型

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + u$$

其中 $u \sim N(0, \sigma^2)$, 当求得参数 w 的最小二乘估计 $\hat{w}$ 之后, 就可以建立回归方程

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 + \cdots + \hat{w}_m x_m$$

而且在经过线性回归显著性及回归系数显著性的检验后, 表明回归方程和回归系数都是显著的, 那么就可以利用回归方程来进行预测。给定自变量 $x_1, x_2, \cdots, x_m$ 的任意一组观察值 $x_{01}, x_{02}, \cdots, x_{0m}$, 由回归方程可得

$$\hat{y}_0 = \hat{w}_0 + \hat{w}_1 x_{01} + \hat{w}_2 x_{02} + \cdots + \hat{w}_m x_{0m}$$

设 $\mathbf{x}_0 = (1, x_{01}, x_{02}, \cdots, x_{0m})$, 则上式可以写成

$$\hat{y}_0 = \mathbf{x}_0 \hat{\mathbf{w}}$$

在 $\mathbf{x} = \mathbf{x}_0$ 时, 由样本回归方程计算的 $\hat{y}_0$ 是个别值 y_0 和总体均值 $E(y_0)$ 的无偏估计, 所以 $\hat{y}_0$ 可以作为 y_0 和 $E(y_0)$ 的预测值。

与第 2 章中讨论的情况相同, 区间预测包括两个方面: 一方面是总体个别值 y_0 的区间预测; 另一方面是总体均值 $E(y_0)$ 的区间预测。设 $e_0 = y_0 - \hat{y}_0 = y_0 - \mathbf{x}_0 \hat{\mathbf{w}}$, 则有

$$e_0 \sim N(0, \sigma^2 [1 + \mathbf{x}_0 (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}_0^T])$$

如果 $\hat{w}$ 是统计模型中某个参数 w 的估计值, 那么 T 统计量的定义式就为

$$t_{\hat{\beta}} = \frac{\hat{w}}{\text{se}(\hat{w})}$$

所以与 e_0 相对应的 T 统计量的表达式如下:

$$T = \frac{e_0}{\hat{\sigma} \sqrt{1 + \mathbf{x}_0 (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}_0^T}} = \frac{y_0 - \hat{y}_0}{\hat{\sigma} \sqrt{1 + \mathbf{x}_0 (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}_0^T}} \sim t(n - m - 1)$$

在给定显著水平 α 的情况下, 可得

$$\begin{aligned} \hat{y}_0 - t_{\frac{\alpha}{2}}(n - m - 1) \times \hat{\sigma} \sqrt{1 + \mathbf{x}_0 (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}_0^T} &\leq y_0 \\ &\leq \hat{y}_0 + t_{\frac{\alpha}{2}}(n - m - 1) \times \hat{\sigma} \sqrt{1 + \mathbf{x}_0 (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}_0^T} \end{aligned}$$

总体个别值 y_0 的区间预测就由上式给出。

在 Python 中如果使用 statsmodels 包来构建回归模型, 那么调用 predict() 方法可以根据新的输入来进行回归预测。针对土壤沉淀物吸收的例子中, 可以使用下面的命令来预测当可提取的铁含量为 150, 可提取的铝含量为 40 时, 磷酸盐的吸收情况为:

```
print(model_multilin.predict([1, 150, 40]))
```

如果希望得到预测结果的区间估计,则需要用到 `get_prediction()` 方法:

```
predictions = model_multilin.get_prediction([1, 150, 40])
predictions.summary_frame(alpha = 0.05)
```

执行上述代码,程序输出如图 3-8 所示,其中被圈出的部分就是区间估计结果。

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	23.51929	1.310799	20.598648	26.439933	13.333718	33.704863

图 3-8 带区间估计的结果预测

其中点预测的结果是 23. 519 29,在 5% 的显著水平下,个别值的区间预测结果是 (13. 333 72, 33. 704 86)。

当然也可以根据公式来尝试手动计算一下这个结果,其中的 T 统计量临界值可以由下面的代码求得:

```
t. isf(0.05/2, 10)
```

所得结果为 2. 228 139。

继续计算,为了得到更精确的误差标准值,使用下面的代码进行计算。Python 中自动输出的结果 4. 379 是我们所计算结果在保留 4 位有效数字后得到的。

```
import math
math.sqrt(sum(model_multilin.resid * model_multilin.resid)/10) # 4.379375
```

另外还可以算得 $\mathbf{x}_0(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}_0^T$ 的值为

$$\begin{bmatrix} 1 \\ 150 \\ 40 \end{bmatrix}^T \times \begin{bmatrix} 0.633138 & -0.003826 & 0.002477 \\ -0.003826 & 0.000046 & -0.000088 \\ 0.002477 & -0.000088 & 0.000265 \end{bmatrix} \times \begin{bmatrix} 1 \\ 150 \\ 40 \end{bmatrix} = 0.08958766$$

然后在 Python 中使用下面的代码计算最终的预测区间,可见结果与前面给出的结果是基本一致的。

```
23.51929 - 2.228139 * math.sqrt(1 + 0.08958766) * 4.379375 # 13.33372
23.51929 + 2.228139 * math.sqrt(1 + 0.08958766) * 4.379375 # 33.70486
```

类似地,还可以得到总体均值 $E(y_0)$ 的区间预测表达式为

$$\hat{y}_0 - t_{\frac{\alpha}{2}}(n-m-1) \times \hat{\sigma} \sqrt{\mathbf{x}_0(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}_0^T} \leq E(y_0) \leq \hat{y}_0 + t_{\frac{\alpha}{2}}(n-m-1) \times \hat{\sigma} \sqrt{\mathbf{x}_0(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}_0^T}$$

前面由 `get_prediction()` 方法得到的输出中, `mean_ci_lower` 和 `mean_ci_upper` 给出的就是总体均值的区间预测结果。而且 y_0 期望的置信区间要比 y_0 的置信区间更窄。

同样,下面的代码给出了包含中间过程的手动计算方法,这与刚刚得到的计算结果是一致的。

```
23.51929 - 2.228139 * math.sqrt(0.08958766) * 4.379375      # 20.59865
23.51929 + 2.228139 * math.sqrt(0.08958766) * 4.379375      # 26.43993
```

3.6 格兰杰因果关系检验

所谓因果关系,可以通过变量之间的依赖性来定义,即作为结果的变量是由作为原因的变量所决定的,原因变量的变化引起结果变量的变化。

因果关系不同于相关关系,而且从一个回归关系式我们并不能确定变量之间是否具有因果关系。虽然我们说回归方程中解释变量是被解释变量的原因,但是,这一因果关系通常是先验设定的,或者是在回归之前就已确定。

实际上,在许多情况下,变量之间的因果关系并不总像农作物产量和降雨量之间的关系那样一目了然,或者没有充分的知识使我们认清变量之间的因果关系。此外,即使某一经济理论宣称某两个变量之间存在一种因果关系,也需要给予经验上的支持。

诺贝尔经济学奖获得者、英国经济学家克莱夫·格兰杰(Clive Granger),是著名的经济时间序列分析大师,被认为是世界上最伟大的计量经济学家之一。格兰杰从预测的角度给出了因果关系的一种描述性定义,这就是我们现在所熟知的格兰杰因果关系。

格兰杰指出,如果一个变量 x 无助于预测另一个变量 y ,则说 x 不是 y 的原因。相反,若 x 是 y 的原因,则必须满足两个条件:第一, x 应该有助于预测 y ,即在 y 关于 y 的过去值的回归中,添加 x 的过去值作为独立变量应当显著地增加回归的解释能力;第二, y 不应当有助于预测 x ,其原因是,如果 x 有助于预测 y , y 也有助于预测 x ,则很可能存在一个或几个其他变量,它们既是引起 x 变化的原因,也是引起 y 变化的原因。现在人们一般把这种从预测的角度定义的因果关系称为格兰杰因果关系。

变量 x 是否为变量 y 的格兰杰原因,是可以检验的。检验 x 是否为引起 y 变化的格兰杰原因的过程如下:

第一步,检验原假设“ $H_0: x$ 不是引起 y 变化的格兰杰原因”。首先,估计下列两个回归模型。

$$\text{无约束回归模型}(u): y_t = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i} + \sum_{i=1}^q \beta_i x_{t-i} + \epsilon_t$$

$$\text{有约束回归模型}(r): y_t = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i} + \epsilon_t$$

式中, α_0 表示常数项; p 和 q 分别为变量 y 和 x 的最大滞后期数,通常可以取的稍大一些; ϵ_t 为白噪声。

然后,用这两个回归模型的残差平方和 RSS_u 与 RSS_r 构造 F 统计量:

$$F = \frac{(\text{RSS}_r - \text{RSS}_u)/q}{\text{RSS}_u/(n - p - q - 1)} \sim F(q, n - p - q - 1)$$

其中, n 为样本容量。

检验原假设“ $H_0: x$ 不是引起 y 变化的格兰杰原因”(等价于检验 $H_0: \beta_1 = \beta_2 = \dots = \beta_q = 0$) 是否成立。如果 $F \geq F_\alpha(q, n - p - q - 1)$, 则 $\beta_1, \beta_2, \dots, \beta_q$ 显著不为 0, 应拒绝原假设“ $H_0: x$ 不是引起 y 变化的格兰杰原因”; 反之, 则不能拒绝原假设。

第二步, 将 y 与 x 的位置交换, 按同样的方法检验原假设“ $H_0: y$ 不是引起 x 变化的格兰杰原因”。

第三步, 要得到“ x 是 y 的格兰杰原因”的结论, 必须同时拒绝原假设“ $H_0: x$ 不是引起 y 变化的格兰杰原因”和接受原假设“ $H_0: y$ 不是引起 x 变化的格兰杰原因”。

最后, 给出一个在 Python 中进行格兰杰因果关系检验的实例。从一般的认识来讲, 消费与经济增长之间存在相互促进的作用。但是, 相比之下, 二者中哪一个对另外一个有更强的促进作用在各国经济发展过程中则呈现出不同的结论。就中国而言, 自改革开放以来, 我们都认为经济增长对消费的促进作用要大于消费对经济增长的促进作用。或者说, 在我国经济增长可以作为消费的格兰杰原因, 反之不成立。但这一结论是否能够得到计量经济研究的支持呢? 下面就在 Python 中通过对 1978—2002 年的统计数据(来源为国家统计局网站)展开分析, 进而来回答这个问题。

首先将 GDP 和消费数据存储在两个数组对象中, 然后再拼接到一起。

```
import numpy as np
import statsmodels.api as sm
from statsmodels.tsa.stattools import grangercausalitytests

GDP = np.array([3645.2, 4062.6, 4545.6, 4889.5, 5330.5, 5985.6, 7243.8, 9040.7,
10274.4, 12050.6, 15036.8, 17000.9, 18718.3, 21826.2, 26937.3, 35260.0,
48108.5, 59810.5, 70142.5, 78060.8, 83024.3, 88479.2, 98000.5, 108068.2, 119095.7])

consumption = np.array([1759.1, 2014, 2336.9, 2627.5, 2867.1, 3220.9, 3689.5,
4627.4, 5293.5, 6047.6, 7532.1, 8778, 9435, 10544.5, 12312.2, 15696.2,
21446.1, 28072.9, 33660.3, 36626.3, 38821.8, 41914.9, 46987.8, 50708.8, 55076.4])

X_gdp_cons = np.vstack((GDP, consumption)).transpose()
X_cons_gdp = np.vstack((consumption, GDP)).transpose()
```

在 Python 中执行格兰杰因果关系检验, 同样可以使用 statsmodels 包。具体来说需要用到的是 grangercausalitytests 类。实例化该类时, 最重要的参数有两个。首先是一个包含有两列时间序列数据的多维数组对象, 后续的格兰杰校验将检验位于第二列的时间序列数据是否是第一列时间序列数据的格兰杰原因。因此, 上述代码中构建了两个输入数据对象, 二者的区别在于 GDP 和消费数据所在的位置不同。另外一个参数用于设置最大滞后期数。如果输入的是一个整数 m , 那么表示要对从 $1 \sim m$ 的各个最大滞后期数都进行一次格兰杰

因果关系校验。如果输入的是 $[m]$,则表示只对 m 这个最大滞后期数做格兰杰因果关系校验。

首先来检验消费是否是 GDP 的格兰杰原因:

```
gc_res_1 = grangercausalitytests(X_gdp_cons, [2])
```

上述代码的输出结果如下:

```
Granger Causality
number of lags (no zero) 2
ssr based F test:          F = 1.6437,          p = 0.2210, df_denom = 18, df_num = 2
ssr based chi2 test:       chi2 = 4.2005,        p = 0.1224, df = 2
likelihood ratio test:     chi2 = 3.8581,        p = 0.1453, df = 2
parameter F test:         F = 1.6437,          p = 0.2210, df_denom = 18, df_num = 2
```

接下来检验 GDP 是否是消费的格兰杰原因:

```
gc_res_2 = grangercausalitytests(X_cons_gdp, [2])
```

上述代码的输出结果如下:

```
Granger Causality
number of lags (no zero) 2
ssr based F test:          F = 13.4108,         p = 0.0003, df_denom = 18, df_num = 2
ssr based chi2 test:       chi2 = 34.2720,        p = 0.0000, df = 2
likelihood ratio test:     chi2 = 20.9833,        p = 0.0000, df = 2
parameter F test:         F = 13.4108,         p = 0.0003, df_denom = 18, df_num = 2
```

检验结果显示,当原假设为“消费不是引起 GDP 变化的格兰杰原因”, F 测试的 P 值结果为 $0.221 > 0.05$,我们无法拒绝原假设;而当原假设变为“GDP 不是引起消费变化的格兰杰原因”时, F 测试的 P 值为 $0.0003 < 0.05$,我们可以据此拒绝原假设。于是可以证明:GDP 是消费的格兰杰原因。

之前计算结果中的 P 值也可以用下面的 Python 代码算得(注意当 $p=2$ 时,观测值的数量 $n=25-2=23$):

```
f.sf(1.6437, 2, 18)      #0.2209779592
f.sf(13.411, 2, 18)      #0.0002716636
```