Python机器学习技术与应用

张爱桃

河北医科大学

中国水利水电出版社
www.waterpub.com.cn

第五章

多元回归分析

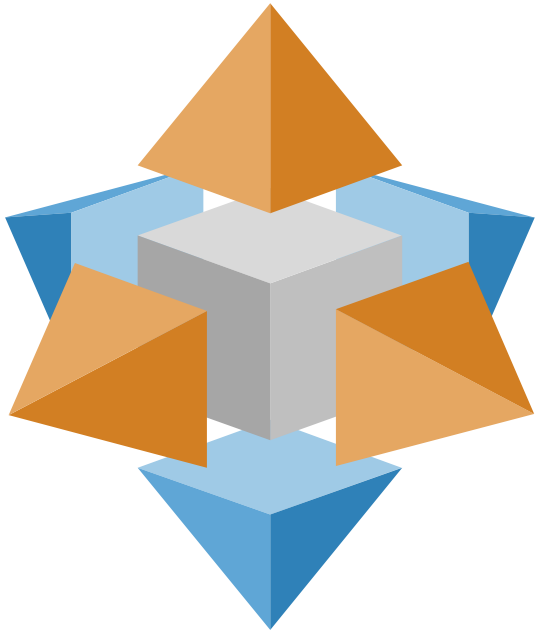
中国水利水电出版社
www.waterpub.com.cn

本章要点

线性回归的基本原理

多元线性回归的实现

回归模型的评估指标



多重共线性

岭回归、Lasso回归

多项式回归、Logistic回归

主要内容

5.1 多元回归分析

5.2 多重共线性问题

5.3非线性回归-多项式回归

5.4 Logistic 回归

5.1 多元回归分析

- **一元线性回归模型：**

因变量y与自变量x的线性回归模型为:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (5.1)$$

其中β_0，β_1是回归模型的参数，称为回归系数（β_0也称为截距项），ε是随机误差。

- **多元线性回归模型：**

因变量y关于k(k>=2)个自变量x_1 x_2,…, x_k的多元线性回归模型为:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (5.2)$$

在回归分析中，对ε有以下几点基本假设：

（1）ε的均值是0，即：E(e)=0

（2）ε对于自变量x_1 x_2,…, x_k的所有取值具有同方差

（3）ε的概率分布服从正态分布即：$\varepsilon \sim N(0, \sigma^2)$，且相互独立

- **估计模型参数：**

假设n组样本数据（$x_{i1}, x_{i2}, \cdots x_{ik}; y_i$），i=1,2,⋯,n，$y_i$是观测值，因变量$y_i$的回归值$\hat{y}_i$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} \tag{5.5}$$

y_i与$\hat{y}_i$的残差为：

$$e_i = (y_i - \hat{y}_i) = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik})] \tag{5.6}$$

所有n组数据，观察值y_i与预测值$\hat{y}_i$残差平方和为

$$\mathrm{SSE} = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik})]^2 \tag{5.7}$$

使SSE取得最小值的$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，⋯，$\hat{\beta}_k$称为参数$b_0, b_1, b_2, \cdots, b_k$**最小二乘估计**。

- **估计模型参数：**

最小二乘方程组：

当k的数值较大时通常将多元线性回归模型表述为矩阵，利用矩阵代数求解线性方程组。

令 $Y=\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$ $X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$ $\hat{\beta}=\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$ $e=\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$

利用这些符号，最小二乘方程组可以写成如下矩阵的方程：

$$(X'X)\hat{\beta}=X'Y \qquad (5.8)$$

方程的解为：

$$\hat{\beta}=(X'X)^{-1}X'Y \qquad (5.9)$$

利用最小二乘法求解时，要求X为满秩矩阵，即各个自变量$x_1, x_2, \cdots, x_k$是不相关的。

5.1 多元回归分析

- 例：某研究机构进行了一系列实验来收集关于混凝土渗透性系数的信息。9组实验的混凝土渗透性系数y与孔隙率x_1，渗水率x_2测量值列于表5.1中，根据表5.1中的数据试建立渗透性系数与其他几项指标关系的多元线性回归方程。

序号	孔隙率 x1	渗水率 x2	渗透性系数 y
1	0.050	0.903	1.00
2	0.035	0.722	1.00
3	0.025	0.590	1.00
4	0.050	0.345	0.10
5	0.035	0.282	0.10
6	0.025	0.233	0.10
7	0.050	0.103	0.01
8	0.035	0.091	0.01
9	0.025	0.078	0.01

Y、X和 $\hat{\beta}$ 的矩阵如下：

$$Y=\begin{bmatrix}1.00\\1.00\\1.00\\0.10\\0.10\\0.10\\0.01\\0.01\\0.01\end{bmatrix}\quad X=\begin{bmatrix}1 & 0.050 & 0.903\\1 & 0.035 & 0.722\\1 & 0.025 & 0.590\\1 & 0.050 & 0.345\\1 & 0.035 & 0.282\\1 & 0.025 & 0.233\\1 & 0.050 & 0.103\\1 & 0.035 & 0.091\\1 & 0.025 & 0.078\end{bmatrix}\quad \hat{\beta}=\begin{bmatrix}\hat{\beta}_0\\\hat{\beta}_1\\\hat{\beta}_2\end{bmatrix}$$

那么

$$\hat{\beta}=\begin{bmatrix}\hat{\beta}_0\\\hat{\beta}_1\\\hat{\beta}_2\end{bmatrix}=(X'X)^{-1}X'Y=\begin{bmatrix}0.1320\\-9.3071\\1.5578\end{bmatrix}$$

因此所求任务得分的预测回归方程为：

$$\hat{y}_i=0.1320-9.3071\,x_{i1}+1.5578x_{i2}$$

- 回归模型的评估指标：

1．平均绝对值误差（Mean absolute error，MAE）

这个指标是对绝对误差损失的预期值。

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|) \qquad (5.10)$$

程序如下

metrics.mean_absolute_error(y_test, pred_y)

2．均方误差（Mean square error ,MSE）

均方误差是观察值与预测值的平方差的平均值，它是模型拟合的绝对度量。

$$MSE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5.11)$$

MSE的值越小，误差越小，回归方程拟合程度越高，模型效果越好。它是线性回归中最常用的损失函数，线性回归过程中尽量让该损失函数最小。

mean_squared_error(y_test, pred_y) # 调用函数实现均方误差(损失值)计算

- 回归模型的评估指标：

3．均方根误差（Root mean square error ,RMSE）

这个指标是对绝对误差损失的预期值。

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \qquad (5.12)$$

4．中值绝对误差（Median absolute error, MedAE）

$$MedAE(y,\hat{y}) = \text{median}(|y_1-\hat{y}_1|,\cdots,|y_n-\hat{y}_n|) \qquad (5.13)$$

该指标对样本量不敏感，即使在比较少的样本数据中依然可行，并且对异常值不敏感，不会因为特殊的异常值而导致估计的严重偏差。此方法非常适用于含有离群点的数据集。

```
metrics.median_absolute_error(y_test, pred_y))  #调用函数实现中值绝对误差计算
```

- 回归模型的评估指标：

5．决定系数（R-square）拟合优度测定

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)}{\sum(y_i - \bar{y})^2} \qquad (5.14)$$

R^2的取值范围为[0,1]，它代表在y值总变异中可由回归模型解释部分所占的百分比，用以反应线性回归模型能在多大程度上解释因变量y的变异性。因此R^2=0 意味着模型对数据完全没有拟合，R^2=1 意味着完美拟合。

R2 = metrics.r2_score(y_test, pred_y) 或者
R2 = clf.score(x_test, y_test)

5.2 多重共线性问题

例：22例胎儿的身长，头围，体重和胎儿的受精周龄的测量数据如表5.4所示，建立由前三个指标推测胎儿受精周龄的回归方程。

序号	身长（cm） x_1	头围（cm） x_2	体重（g） x_3	受精周龄 y
1	13.00	9.20	50.00	13.00
2	18.70	13.20	102.00	14.00
3	21.00	14.80	150.00	15.00
4	19.00	13.30	110.00	16.00
5	22.80	16.00	200.00	17.00
6	26.00	18.20	330.00	18.00
7	28.00	19.70	450.00	19.00
8	31.40	22.50	450.00	20.00
9	30.30	21.40	550.00	21.00
10	29.20	20.50	640.00	22.00
11	36.20	25.20	800.00	23.00
12	37.00	26.10	1090.00	24.00
13	37.90	27.20	1140.00	25.00
14	41.60	30.00	1500.00	26.00
15	38.20	27.10	1180.00	27.00
16	39.40	27.40	1320.00	28.00
17	39.20	27.60	1400.00	29.00
18	42.00	29.40	1600.00	30.00
19	43.00	30.00	1600.00	31.00
20	41.10	27.20	1400.00	33.00
21	43.00	31.00	2050.00	35.00
22	49.00	34.80	2500.00	36.00

拟合的回归方程为

$$y = 11.0117 + 1.6927x_1 - 2.1588x_2 + 0.075x_3$$

- 岭回归（Ridge estimate）

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'Y \qquad (5.15)$$

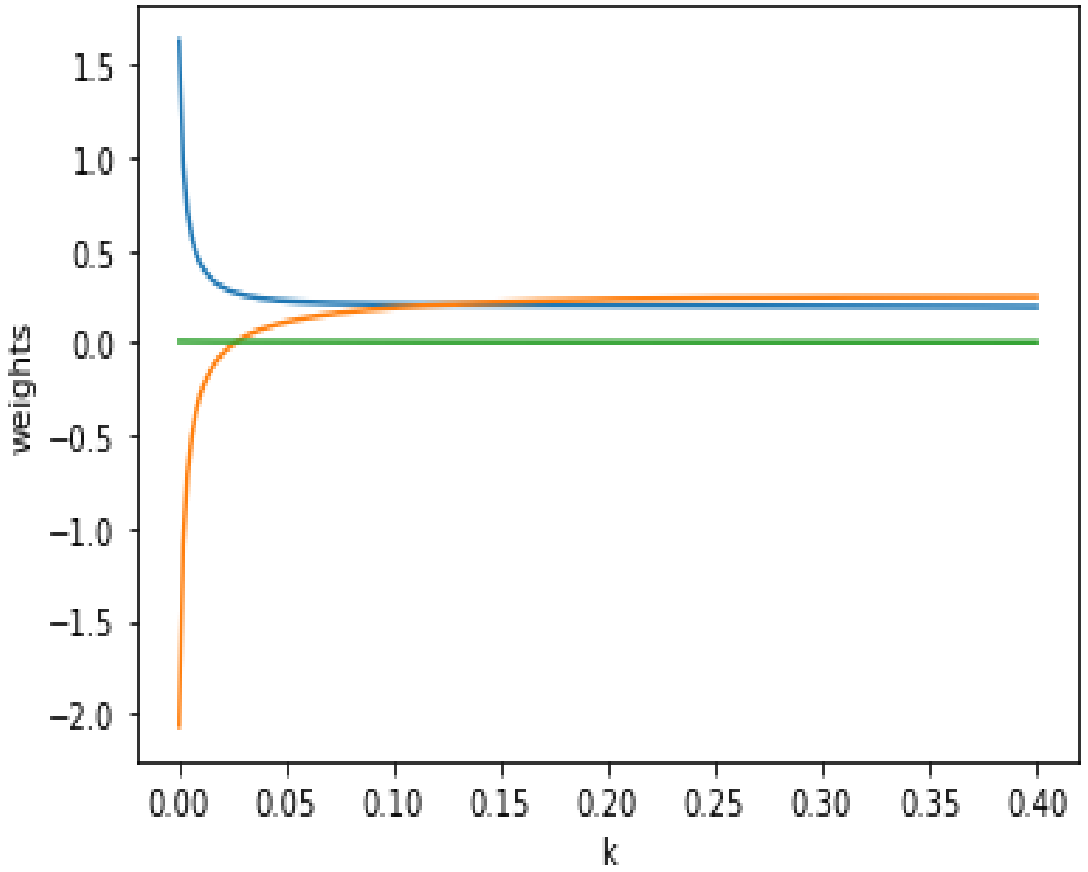
其中k为岭参数，k=0时的岭回归估计就是最小二乘估计，I是单位矩阵。kI项的加入使得$X'X + kI$满秩，保证了可逆，但是也使得回归系数β的估计不再是无偏估计。

岭回归估计的目标函数（观测值和预测值的残差平方和）为 $\mathrm{SSE} = \sum(\mathrm{Y} - X\beta)^2 + \mathrm{k}\|\beta\|^2$。

5.2 多重共线性问题

- K值的选择

k	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0.01	0.42397	-0.27435	0.00653
0.02	0.29810	-0.05807	0.00611
0.03	0.25484	0.03116	0.00579
0.04	0.23463	0.08174	0.00554
0.10	0.20693	0.19149	0.00462
0.20	0.20139	0.23463	0.00396
0.40	0.19376	0.24866	0.00339

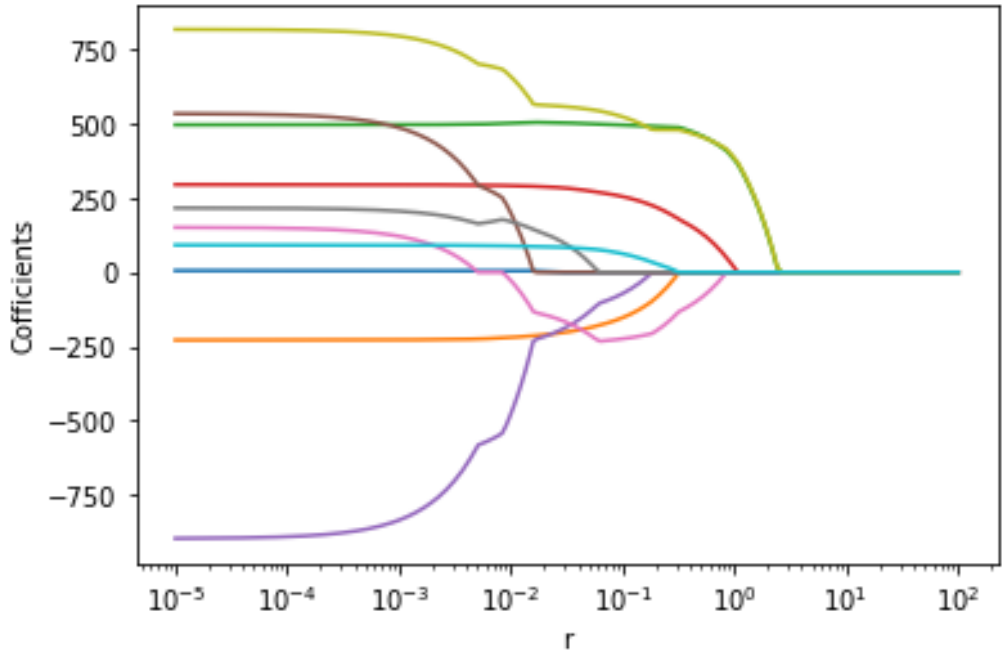


- Lasso回归

Lasso回归的目标函数（损失函数）表达式如下：

$$\mathrm{SSE} = \sum (\mathrm{Y} - X\beta)^2 + \gamma|\beta| \qquad (5.22)$$

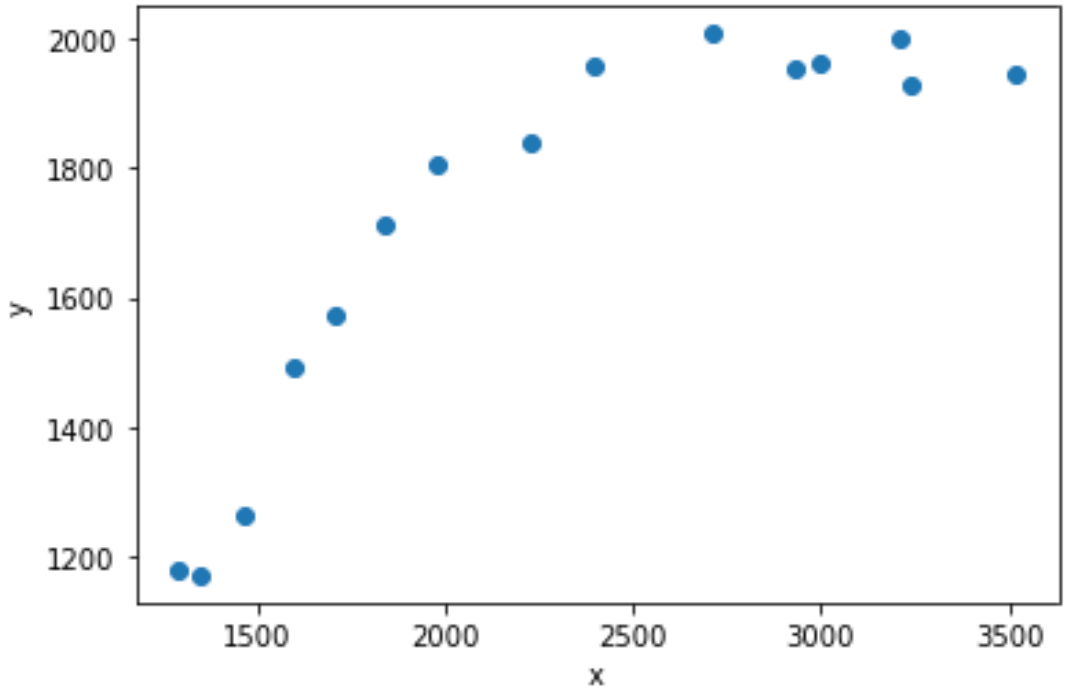
其中，γ为Lasso系数，$|\beta|$为β的L1范数。增加了对回归参数β的L1范数约束。在缩减过程中，随着γ的增大，可以将一些不重要的回归系数直接缩减至0



5.3 多项式回归

- 非线性回归——多项式回归（Polynomial regression）

如某机构调查了全电气化住宅中7月份用电量y和住宅面积x之间的关系，用电量与住宅面积的散点图如图5-4所示。



5.3 多项式回归

- 非线性回归——多项式回归（Polynomial regression）

一元k次多项式回归模型的一般形式为：

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon \qquad (5.16)$$

令 $x_1 = x, x_2 = x^2, \cdots x_k = x^k$ 则

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

因此多项式回归模型就是多元线性回归模型的一个特例

两个或两个以上的自变量的情况，二元二次多项式的回归模型为：

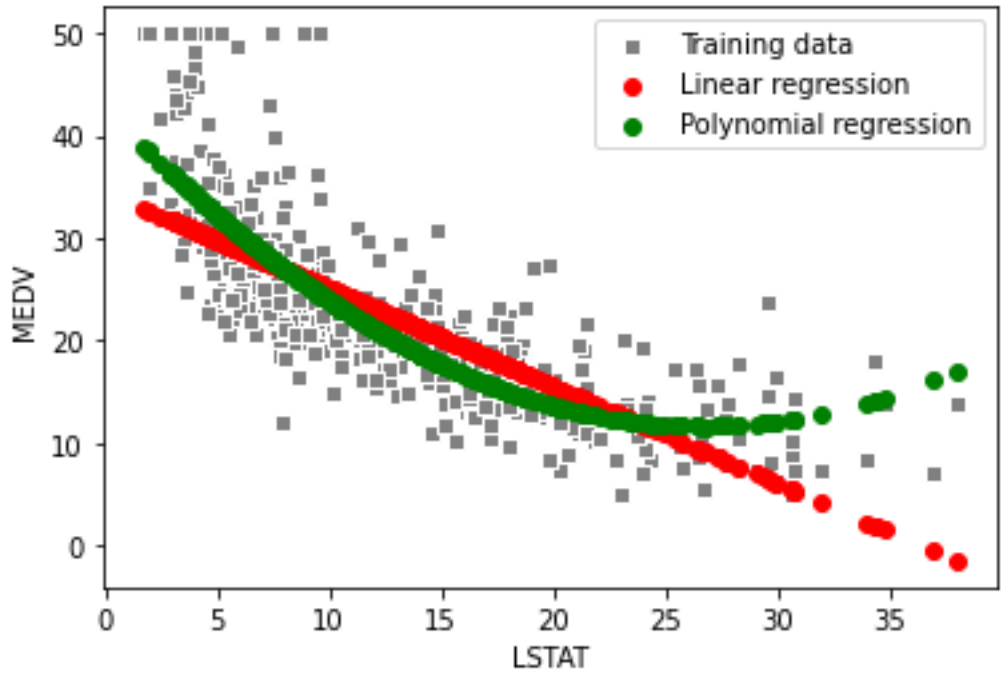
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_1}^2 + \beta_4 {x_2}^2 + \beta_5 x_1 x_2 + \varepsilon \qquad (5.17)$$

其中y为因变量，x_1, x_2为自变量，$x_1 x_2$ 是x_1, x_2的交叉乘积项，表示x_1, x_2的交互作用。系数β_5称为交互影响系数。

5.3 多项式回归

- 非线性回归——多项式回归（Polynomial regression）

分析sklearn自带的波士顿房价预测数据集中人口中低收入阶层比例（特征LSTAT）与房价（标记MEDV）的关系。



5.4 Logistic回归

- Logistic回归模型

是一种广义的线性回归分析模型，是在用线性回归的方法做分类任务。

例如，临床中研究胃癌的影响因素，自变量为年龄、性别、饮食习惯、幽门螺杆菌感染等。因变量为是否胃癌，值为“是”或“否”，即为定性的分类变量。

（1）因变量y本身只有“1”和“0”两个离散值
（2）因变量y的最大值为1，最小值为0

因此需要将线性回归方程的输出转化到{0，1}上

5.4 Logistic回归

阶跃函数的形式为：

$$sgn(z)=\begin{cases}1, & z\geq 0\\ 0, & z<0\end{cases} \qquad (5.18)$$

令$z=\beta_0+\beta_1 x_{i1}+\beta_2 xi_2+\cdots+\beta_k x_{ik}+\varepsilon$,

$$y_i=\begin{cases}1, & z=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon\geq 0\\ 0, & z=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon<0\end{cases} \qquad (5.19)$$

单位阶跃函数并非一个连续函数，不符合线性回归模型的假定，因此常用Logisitc函数来近似。
Logisitc函数的形式为：

$$f(z)=\frac{e^z}{1+e^z} \qquad (5.20)$$

e为欧拉常数，该函数同样能够将输入范围$(-\infty,+\infty)$映射到(0,1)之间

5.4 Logistic回归

近似的回归方程为：

$$y_i = \frac{e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon)}}{1+e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon)}} \qquad (5.21)$$

由于y_i是0-1型变量，$E(y_i)=P(y_i=1)=p_i$是自变量为x_i时，$y_i=1$的概率。$\hat{y}_i$为$E(y_i)$的一个估计，所以我们可以用$y_i=1$的概率p_i代替y_i作为因变量，从而得到Logisitc回归模型为：

$$p_i = \frac{e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon)}}{1+e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon)}}, \qquad i=1,2,3\cdots n \qquad (5.22)$$

对p_i进行如下Logit变换：

$$\acute{p}_i = ln\frac{p_i}{1-p_i} \qquad (5.23)$$

变换后的Logisitc回归模型如下：

$$\acute{p}_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon \qquad (5.24)$$

变换后的$\acute{p}_i$与自变量x_i之间存在线性关系。

5.4 Logistic回归

例某银行信用卡中心拟研究某卡种的持卡人收入水平与违约风险的关系，违约的持卡人记为1，没有违约的持卡人记为0。以持卡人的月收入为自变量x对表5.5所示的数据，建立Logistic回归的模型。

序号	月收入（万元） x_i	持卡人数 n_i	人数 m_i	比 $p_i=\frac{m_i}{n_i}$	No Image
1	0.4	370	74	0.2	-1.386
2	0.6	300	54	0.18	-1.52
3	0.8	320	48	0.15	-1.73
4	1.0	450	45	0.1	-2.2
5	1.2	400	24	0.06	-2.75
6	1.4	520	26	0.05	-2.94
7	1.6	350	14	0.04	-3.18
8	1.8	450	9	0.02	-3.89

因此每组的数据（x_i，$\acute{p}_i$）的Logistic回归方程为 $\acute{p}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$ (5.25)

计算出回归参数为：

$$\hat{\beta}_0=-0.484$$
$$\hat{\beta}_1=-1.788$$

可得回归方程为： $\acute{p}=-0.484-1.788x$

经计算该回归模型的决定系数R^2=0.972，F统计检验量显著性检验P值≈0，高度显著。因此可以应用该Logistic回归方程对违约率做预测。

5.4 Logistic回归

例某银行信用卡中心拟研究某卡种的持卡人收入水平与违约风险的关系，违约的持卡人记为1，没有违约的持卡人记为0。以持卡人的月收入为自变量x对表5.5所示的数据，建立Logistic回归的模型。

序号	月收入（万元） x_i	持卡人数 n_i	人数 m_i	比 $p_i=\frac{m_i}{n_i}$	No Image
1	0.4	370	74	0.2	-1.386
2	0.6	300	54	0.18	-1.52
3	0.8	320	48	0.15	-1.73
4	1.0	450	45	0.1	-2.2
5	1.2	400	24	0.06	-2.75
6	1.4	520	26	0.05	-2.94
7	1.6	350	14	0.04	-3.18
8	1.8	450	9	0.02	-3.89

因此每组的数据（x_i，$\acute{p}_i$）的Logistic回归方程为 $\acute{p}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$ (5.25)

计算出回归参数为：

$$\hat{\beta}_0=-0.484$$
$$\hat{\beta}_1=-1.788$$

可得回归方程为： $\acute{p}=-0.484-1.788x$

经计算该回归模型的决定系数R^2=0.972，F统计检验量显著性检验P值≈0，高度显著。因此可以应用该Logistic回归方程对违约率做预测。

5.4 Logistic回归

Logistic回归评价指标

对于0-1变量的二分类问题，分类的最终结果可以以下矩阵来表示：

预测值	实际值	
	1	0
1	真阳性TP	假阳性FP
0	假阴性FN	真阴性TN

通常将上述矩阵称为“混淆矩阵”。一般情况下，因变量取值为 1 时，将其称为正例（Positive），取值为 0 时称为负例（Negative）。

其中：

TP 表示正确预测正例的样本个数

FP 表示预测为正例但实际为负例的样本个数

FN 表示预测为负例但实际为正例的样本个数

TN 表示正确预测负例的样本个数