多元线性回归分析 Linnerud 数据集

190000000　电气类搬砖　猫九

2021 年 12 月 10 日

1　多元线性回归模型

多元线性回归分析的模型为

$$\begin{cases} y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\varepsilon \\ \varepsilon \sim N\left(0, \sigma^2\right) \end{cases}$$

式中: $\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$ 都是与 $x_1, x_2, \cdots, x_m$ 无关的末知参数, $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$ 称为回归系数。现得到 n 个独立观测数据 $[b_i, a_{i1}, \cdots, a_{im}]$, 其中 b_i 为 y 的观察值, $a_{i1}, \cdots, a_{im}$ 分别为 $x_1, x_2, \cdots, x_m$ 的观察值, $i=1, \cdots, n, n>m$, 由式得

$$\begin{cases} b_i=\beta_0+\beta_1 a_{i1}+\cdots+\beta_m a_{im}+\varepsilon_i \\ \varepsilon_i \sim N\left(0, \sigma^2\right), \quad i=1, \cdots, n \end{cases}$$

记

$$\boldsymbol{X}=\begin{bmatrix} 1 & a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & a_{n1} & \cdots & a_{nm} \end{bmatrix}, \quad \boldsymbol{Y}=\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

$$\boldsymbol{\varepsilon}=\left[\varepsilon_1, \cdots, \varepsilon_n\right]^{\mathrm{T}}, \boldsymbol{\beta}=\left[\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m\right]^{\mathrm{T}}$$

表示为

$$\begin{cases} \boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 \boldsymbol{E}_n\right) \end{cases}$$

式中: $\boldsymbol{E}_n$ 为 n 阶单位矩阵。

2 参数估计

上面中的参数 $\beta_0,\beta_1,\cdots,\beta_m$ 用最小二乘法估计, 即应选取估计值 $\hat{\beta}_j$, 使当 $\beta_j=\dot{\beta}_j, j=0,1,\cdots,m$ 时, 误差平方和

$$Q=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(b_i-\hat{b}_i\right)^2=\sum_{i=1}^{n}\left(b_i-\beta_0-\beta_1 a_{i1}-\cdots-\beta_m a_{im}\right)^2$$

达到最小。为此, 令

$$\frac{\partial Q}{\partial \beta_j}=0, j=0,1,2,\cdots,n$$

得

$$\begin{aligned}
&\frac{\partial Q}{\partial \beta_0}=-2\sum_{i=1}^{n}\left(b_i-\beta_0-\beta_1 a_{i1}-\cdots-\beta_m a_{im}\right)=0\\
&\frac{\partial Q}{\partial \beta_j}=-2\sum_{i=1}^{n}\left(b_i-\beta_0-\beta_1 a_{i1}-\cdots-\beta_m a_{im}\right)a_{ij}=0,\quad j=1,2,\cdots,m。
\end{aligned}$$

经整理化为以下方程组:

$$\begin{gathered}
\beta_0 n+\beta_1\sum_{i=1}^{n}a_{i1}+\beta_2\sum_{i=1}^{n}a_{i2}+\cdots+\beta_m\sum_{i=1}^{n}a_{im}=\sum_{i=1}^{n}b_i\\
\beta_0\sum_{i=1}^{n}a_{i1}+\beta_1\sum_{i=1}^{n}a_{i1}^2+\beta_2\sum_{i=1}^{n}a_{i1}a_i+\cdots+\beta_m\sum_{i=1}^{n}a_{i1}a_{im}=\sum_{i=1}^{n}a_{i1}b_i\\
\vdots\\
\beta_0\sum_{i=1}^{n}a_{im}+\beta_1\sum_{i=1}^{n}a_{im}a_{i1}+\beta_2\sum_{i=1}^{n}a_{im}a_{i2}+\cdots+\beta_m\sum_{i=1}^{n}a_{im}^2=\sum_{i=1}^{n}a_{im}b_i
\end{gathered}$$

方程组的矩阵形式为

$$\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{X}^{\top}\boldsymbol{Y}$$

当矩阵 $\boldsymbol{X}$ 列满秩时, $\boldsymbol{X}^{\top}\boldsymbol{X}$ 为可逆方阵, 解为

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

将 $\boldsymbol{\beta}$ 代回原模型得到 y 的估计值, 即

$$\hat{y}=\hat{\beta}_0+\dot{\beta}_1 x_1+\cdots+\dot{\beta}_m x_m$$

而这组数据的拟合值为

$$\hat{b}_i=\hat{\beta}_0+\hat{\beta}_1 a_{i1}+\cdots+\hat{\beta}_m a_{im}(i=1,\cdots,n)$$

记 $\hat{\boldsymbol{Y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}=\left[\hat{b}_1,\cdots,\hat{b}_n\right]^{\mathrm{T}}$, 拟合误差 $\boldsymbol{e}=\boldsymbol{Y}-\hat{\boldsymbol{Y}}$ 称为残差, 可作为随机误差 $\boldsymbol{\varepsilon}$ 的估计, 而

$$Q=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(b_i-b_i\right)^2$$

为残差平方和。

3 回归分析

本例中采用 Linnerud 给出的关于体能训练的数据进行多元线性回归建模。数据集中被测的样本点是某健身俱乐部的 20 名中年男子, 被测变量分为两种, 第一组是身体特征指标包括体重, 腰围和脉搏。第二组变量是训练结果指标, 包括单杠, 弯曲和跳高。

令 x_1,x_2,x_3 分别表示自变量指标 **weight**、**waist**、**pulse**，y_1，y_2，y_3 分别表示因变量指标 **chins** 、**situps** 、**jumps**。

求 y_1,y_2,y_3 关于 x_1,x_2,x_3 的线性回归方程，

$$y_1=c_{10}+c_{11}x_1+c_{12}x_2+c_{13}x_3$$

$$y_2=c_{20}+c_{21}x_1+c_{22}x_2+c_{23}x_3$$

$$y_3=c_{30}+c_{31}x_1+c_{32}x_2+c_{33}x_3$$

计算 c 矩阵的估计值，经过计算可知：

$$\boldsymbol{c_{ij}}=\begin{bmatrix}221.1277 & 1.5085 & -0.3932 & 0.0789\\ 41.1290 & -0.0638 & -0.0470 & -0.0470\\ 49.4081 & -0.2045 & 0.0733 & -0.0420\end{bmatrix}$$

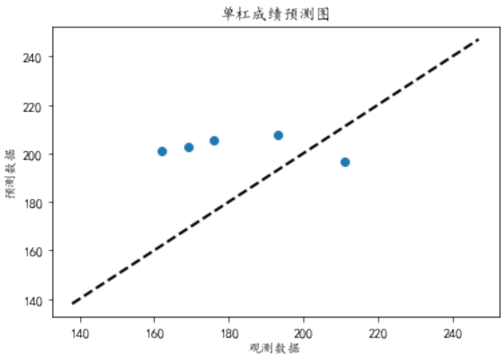
对于多元线性回归模型的质量，可以采用均方根误差 (**Root Mean Squared Error,RMSE**) 对模型进行评价，$RMSE$ 的矩阵为：

$$\boldsymbol{R}=\begin{bmatrix}28.04 & 1.30 & 5.68\end{bmatrix}^T$$

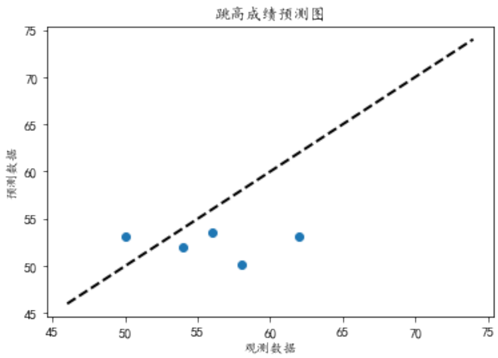
这里用图 1观察真实值与预测值的变化关系，离中间的直线 $y=x$ 越近的点代表预测损失越低。

由上面的回归分析图可知, 对于因变量指标 y_2(弯曲) 预测值最接近真实值, 误差最小,y_2 关于 x_1,x_2,x_3 的回归方程为:

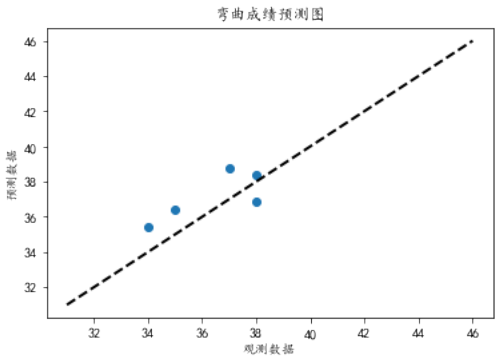
$$y_2=41.1290-0.0638x_1-0.0470x_2-0.0470x_3$$



(a) 单杠成绩预测图



(b) 调高成绩预测图



(c) 弯曲成绩预测图

图 1: 因变量观测值与实际值

对于因变量指标 y_1, y_3 预测值与真实值误差较大, 需要对回归系数进行假设检验和区间估计。

附录

A 代码

```
import pandas as pd
from sklearn.cross_decomposition import PLSRegression
from sklearn.linear_model import LinearRegression
from sklearn import datasets
from sklearn.model_selection import GridSearchCV
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn import metrics
from matplotlib import pyplot as plt

dataset = datasets.load_linnerud()
col_names = dataset['feature_names'] + dataset['target_names']
data = pd.DataFrame(data= np.c_[dataset['data'], dataset['target']],
                                   columns=col_names)
x=np.array(data.loc[:,dataset['feature_names']])
y=np.array(data.loc[:,dataset['target_names'][2]])

x_train,x_test,y_train,y_test = train_test_split(x,y,random_state=1)
linreg = LinearRegression()
linreg.fit(x_train,y_train)

print(linreg.coef_,linreg.intercept_)
y_pred = linreg.predict(x_test)
print(metrics.mean_squared_error(y_test,y_pred))
print(np.sqrt(metrics.mean_squared_error(y_test,y_pred)))

plt.scatter(y_test,y_pred)
plt.plot([y.min(),y.max()],[y.min(),y.max()],'k--',lw=2)
plt.xlabel('Measured')
plt.ylabel('Predicted')
plt.show()
```

B Linnerud 数据集

表 1: Linnerud 数据集

Weight	**Waist**	**Pulse**	**Chins**	**Situps**	**Jumps**
191	36	50	5	162	60
189	37	52	2	110	60
193	38	58	12	101	101
162	35	62	12	105	37
189	35	46	13	155	58
182	36	56	4	101	42
211	38	56	8	101	38
167	34	60	6	125	40
176	31	74	15	200	40
154	33	56	17	251	250
169	34	50	17	120	38
166	33	52	13	210	115
154	34	64	14	215	105
247	46	50	1	50	50
193	36	46	6	70	31
202	37	62	12	210	120
176	37	54	4	60	25
157	32	52	11	230	80
156	33	54	15	225	73
138	33	68	2	110	43