线性回归算法

目录

线性回归研究多 X 对于 Y 的影响关系情况，其位于 SPSSAU-》通用方法-》线性回归。在 SPSSAU 中支持：

- ✓ 拟合线性回归，并且输出各项指标，包括 VIF/DW/RMSE/AIC/BIC 等；
- ✓ 支持保存预测值和残差值；
- ✓ 支持异常值诊断和保存异常点信息值（仅支持<1000 个样本时）；
- ✓ 支持做数据预测。

将分析项拖拽至右侧框然后‘开始分析’。SPSSAU 中涉及三项参数，如下：

- ✓ 保存残差和预测值：选中该参数后，SPSSAU会将残差和预测值保存为新的标题，标题名称类似为“Regression_Residual_****”和“Regression_Prediction_****”。
- ✓ 异常点信息保存：选中该参数后，SPSSAU会将Leverage、Cook和Sres保存为新的标题，标题名称类似为“Regression_Leverage_****”、“Regression_Cook_****”和“Regression_Sres_****”。
- ✓ 异常点诊断：选中该参数后，SPSSAU会输出异常点分析结果。

如果需要使用逐步线性回归，可使用 SPSSAU 进阶方法模块》逐步回归，如果回归时带‘加权’项，可使用计量研究模块里面的 OLS 回归。除此之外，SPSSAU 进阶方法模块有提供分层回归，计量研究模块有提供分组回归，其数学原理均为线性回归。

SPSSAU 数据格式

Y	X1	X2	X3	X4
3.15	8.49	1.04	5.97	2.991
1.729	9.04	6.355	0.056	0.685
2.772	7.395	0.22	5.161	4.192
0.097	7.635	5.999	5.071	9.81
3.399	3.215	7.433	6.758	6.132
6.829	9.652	0.974	5.025	0.801
8.85	4.582	4.263	1.61	2.442
3.816	9.396	5.577	2.442	4.42
5.655	9.547	1.236	2.765	2.521
2.832	3.771	0.961	1.08	7.885
0.379	5.6	1.172	3.653	2.169
1.227	0.774	6.439	9.044	6.757
2.789	6.835	1.825	1.938	4.818
0.517	3.258	6.731	7.311	9.282
3.953	5.687	7.187	9.588	6.094
3.486	0.603	4.138	5.393	6.426
9.39	6.1	9.759	5.484	7.593
5.408	0.13	6.973	1.052	1.954
7.127	7.114	6.621	4.831	9.089

比如上图为研究 3 个 X 对于 Y 的影响关系情况的数据格式。

计算公式

1. 线性回归

假设有p个x，n个样本数据$(x_{i1}, x_{i2}, \dots, x_{ip}, y_i)$,其中$i = 1,2, \dots, n$和$j = 1,2, \dots, p$。线性回归模型可以表示为:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

其中:

β_0: 截距项

$\beta_1, \beta_2, \dots, \beta_p$: 回归系数

ϵ_i: 随机误差项

多元线性回归的参数$\beta_0, \beta_1, \dots, \beta_p$，通过最小化残差平方和来估计:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

首先构建数据矩阵:

$$\boldsymbol{X}=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1p}\\ 1 & x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{np}\end{bmatrix}$$

$$\mathbf{y}=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$$

求解线性方程组

$$(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}=\mathbf{X}^T\mathbf{y}$$

得到参数估计

$$\widehat{\boldsymbol{\beta}}=\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_p\end{bmatrix}$$

得到回归方程

$$\hat{y}=\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j\,x_j$$

另外标准化回归系数

$$\beta_j^*=\frac{\beta_j\ \times\ S_{x_j}}{S_y}$$

其中：

S_{x_j}是自变量x_j的标准差

S_y是因变量y的标准差

针对回归系数$\hat{\beta}_j$，其标准误差$SE(\hat{\beta}_j)$的计算公式：

$$SE(\hat{\beta}_j)=\sqrt{|(\mathbf{X}^T\mathbf{X})^{-1}{}_{ii}|\times MSE}$$

其中：

$(\mathbf{X}^T\mathbf{X})^{-1}{}_{ii}$表示矩阵主对角线数据

计算残差和预测值

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

提示：

选中‘保存残差和预测值’参数后，SPSSAU 会将残差和预测值保存为新的标题。

2. 拟合指标

关于方差分析 ANOVA:

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

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

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

$$SST = SSR + SSE$$

$$MSR = \frac{SSR}{p}$$

$$MSE = \frac{SSE}{n - p - 1}$$

$$F = \frac{MSR}{MSE}$$

$$df_1 = p$$

$$df_2 = n - p - 1$$

其中:

y_i是第i个数据真实y值

$\bar{y}$是y的平均值

$\hat{y}$是预测值

n是样本量

p是模型x的数量

e_i是第i个数据的残差

模型标准误差:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - p - 1}}$$

可决系数:

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

$$R = \sqrt{R^2}$$

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

其它指标:

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

$$DW = \frac{\sum_{i=1}^{n-1}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

$$AIC = nln(SSR) + 2(p+1) - nln(n)$$

$$BIC = nln(SSR) + (p+1)ln(n) - n\ln(n)$$

$$VIF_j = \frac{1}{1 - R_j^2}$$

$$Tolerance_j = \frac{1}{VIF_j} = 1 - R_j^2$$

其中：

R_j^2是自变量x_j作为 y，并且其它x作为自变量时进行线性回归的可决系数。

3. 异常值

SPSSAU 中共提供 3 个异常值判断指标，分别是杠杆值（Leverage）、Cook 距离（Cook）和学生化残差（Sres），该 3 个指标可分别用于异常点的分析和判断，其计算如下：

杠杆值：

$$\mathrm{H} = X(X^T X)^{-1} X^T$$

$$\text{Leverage} = \mathrm{H}_{ii}$$

提示：H_{ii}为H矩阵主对角线数据

学生化残差：

$$Sres_i = \frac{e_i}{\sqrt{MSE \times (1 - \mathrm{H}_{ii})}}$$

Cook 距离：

$$D_i = \frac{1}{p+1}\left(\frac{\mathrm{H}_{ii}}{1 - \mathrm{H}_{ii}}\right) * Sres_i^2$$

提示：

离群点：学生化残差$Sres$的绝对值>2

强影响点：$D_i > \frac{4}{n-p-1}$

强杠杆点：杠杆值（Leverage）$> \frac{3 \times (p+1)}{n}$

选中‘异常值诊断’时，SPSSAU 会输出上述 3 项指标的分析结果和分析建议，如果研究者希望过滤掉异常点，可选选中‘异常点信息保存’，接着分析时按照条件进行‘筛选样本’，使用筛选后的数据进行分析。

参考文献

【1】The SPSSAU project (2024). SPSSAU. (Version 24.0) [Online Application Software]. Retrieved from https://www.spssau.com.

【2】周俊,马世澎. SPSSAU 科研数据分析方法与应用.第 1 版[M]. 电子工业出版社,2024.