回归分析

一元线性回归模型

1.一元线性回归分析

形如

$$y = \beta_0 + \beta_1 x$$

在最小二乘法下，

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

拟合度检验，相关系数：

$$r = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}}$$

$|r| \le 1$， $|r|$ 越大，x，y 线性关系越强.

2.Python求解

调用格式：

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

a = np.loadtxt('data.txt')
re = sm.formula.ols('y ~ x', a).fit()  # 拟合线性回归模型
re.summary()  # 用于查看回归结果的汇总信息的方法
re.get_prediction(a)  # 预测数据
```

多元线性回归模型

1.多元线性回归理论

形如：

$$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m,$$

记数据集

$$X=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1m}\\ 1 & x_{21} & x_{22} & \cdots & x_{2m}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm}\end{bmatrix},\quad Y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix},$$

在最小二乘方法下，

$$\hat{\beta}=(X^TX)^{-1}X^TY,$$

将

$$\hat{\beta}=[b_0,\ b_1,\ \cdots,\ b_m],$$

代入上述回归模型，有

$$y=b_0+b_1x_1+\cdots+b_mx_m.$$

拟合度检验，复相关系数

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

R 越大，模型拟合越好，一般 R 大于 0.8 才认为相关关系成立.

2.Python求解

调用格式：

```
import statsmodels.api as sm

# 基于公式构建并拟合模型
sm.formula.ols(formula, data=df).fit()
# 基于数组构建并拟合模型
sm.OLS(y, X).fit()
```

多项式回归

形如

$$y=\beta_0+\beta_1 x+\cdots+\beta_n x^n,$$

利用python求解:

```
import statsmodels.api as sm

re = sm.formula.ols(formula, data=df).fit()
```

广义线性回归模型

1.分组数据的Logistic回归模型

Logistic函数:

$$f(x)=\frac{1}{1+e^{-x}}$$

在拟合时，通常写为

$$p_i=\frac{1}{1+e^{\beta_0+\beta_1 x_i}},\quad i=1,2,\cdots,c,$$

c 为分组数据的组数，p_i 为样本比例，对上述方程进行 $Logit$ 变换，令

$$p_i^*=ln(\frac{p_i}{1-p_i}),$$

得到

$$p^*=\beta_0+\beta_1 x,$$

按照一般线性回归求解系数即可.
求得回归方程的含义为 在自变量 x_i 的条件下 y_i 等于 1 的比例.

适用于样本量大的分组数据，以组数为回归拟合的样本量，拟合精度低.

2.未分组数据的Logistic回归模型

设 y 为 $0-1$ 型变量，x_1，x_2 $\cdots$，x_m 是与 y 相关的确定性变量，n 组观测数据为

$$(x_{i1},\ x_{i2},\ \cdots,\ x_{im};\ y_i),\quad i=1,2,\ldots,m,$$

满足

$$y_i=f(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_m x_{im}),$$

其中函数 $f(x)$ 为值域在 $[0\ 1]$ 区间的单调增函数.
对应的 $Logistic$ 回归为

$$y_i=\frac{1}{1+e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_mx_{im}}},$$

拟合时，使似然函数的自然对数

$$lnL=\sum_{i=1}^{n}[y_i(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_mx_{im})]-ln(1+e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_mx_{im}})$$

达到极大即可.

python调用：

```
import statsmodels.api as sm

md = sm.formula.glm(formula, data=df, family=sm.families.Binomial()).fit()
```

3.Probit回归模型

回归函数为

$$\phi^{-1}(y_i)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_mx_{im},$$

其中 $\phi(x)$ 为标准正态分布函数.
通常对数据进行 $Probit$ 变换，即

$$p_i^*=\phi^{-1}(p_i),$$

其中 p_i 为样本比例，得一般线性回归方程

$$p_i^*=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_mx_{im}.$$