回归分析

> 具体地说，回归分析是在一组数据的基础上研究这样几个问题：
> (1) 建立因变量y与自变量$x_1, x_2, \cdots, x_m$之间的回归模型（经验公式）；
> (2) 对回归模型的可信度进行检验；
> (3) 判断每个自变量$x_i, (i = 1, 2, \ldots, m)$对$y$的影响是否显著；
> (4) 诊断回归模型是否适合这组数据；
> (5) 利用回归模型对y进行预报或控制

一、一元线性回归模型

1. 模型建立与求解

>> 一般形式

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

其中，β_0和β_1是未知待定系数，$\varepsilon \sim N\left(0, \sigma^2\right)$表示其他随机因素对$y$的影响。

x称为回归变量，y称为响应变量，β_0和β_1称为回归系数。

>> 最小二乘估计

- 偏差平方和

$$Q = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

- 正规方程组

$$\begin{cases} n\beta_0 + \left(\sum\limits_{i=1}^{n} x_i\right)\beta_1 = \sum\limits_{i=1}^{n} y_i \\ \left(\sum\limits_{i=1}^{n} x_i\right)\beta_0 + \left(\sum\limits_{i=1}^{n} x_i^2\right)\beta_1 = \sum\limits_{i=1}^{n} x_i y_i \end{cases}$$

- 参数估计值

$$\hat{\beta}_1 = \frac{L_{xy}}{L_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\text{其中，} L_{xy} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right), L_{xx} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

- 线性回归方程

$$\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

2. 模型分析

>> 一些参数的定义

- 原始数据y_i的变异程度

$$s_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

- 拟合值$\hat{y}_i$的变异程度

$$s_2^2=\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

- 正交分解式

$$\begin{aligned}&\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\\&\Rightarrow\qquad SST=SSR+SSE\end{aligned}$$

- **总变异平方和**：$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2=L_{yy}$

  自由度：$df_T=n-1$
- **可解释的变异平方和**：$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

  自由度：$df_R=1$
- **残差平方和**：$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

  自由度：$df_E=n-2$

>> 相关性检验

> 正交分解式可同时从两个方面说明拟合方程的优良程度：
> (1) SSR越大，用回归方程来解释y_i变异的部分越大，回归方程对原数据解释得越好；
> (2) SSE越小，观测值y_i绕回归直线越紧密，回归方程对原数据的拟合效果越好。

定义**判定系数**（也叫做拟合优度）：

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$$

- $0\le R^2\le 1$，并且R^2越大，拟合效果越好；
- 当$R^2=1$时，有$SSR=SST$，此时原数据的总变异完全可以由拟合值的变异来解释；并且$SSE=0$，即拟合点与原数据完全吻合；
- 当$R^2=0$时，有$SSE=SST$，此时回归方程完全不能解释原数据的总变异，y的变异完全由与x无关的因素引起

定义**相关系数**：

$$r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\pm\sqrt{R^2}$$

- $0\le|r|\le 1$，并且$|r|$越大，线性关系越好
- $r=1$表示正线性相关，$r=-1$表示负线性相关

定义**剩余标准差**：

$$s=\sqrt{\frac{SSE}{n-m}}$$

- 其中，n为样本容量，m为拟合参数个数

>> 显著性检验

提出假设$H_0:\beta_1=0, H_1:\beta_1\neq 0$

若假设$H_0:\beta_1=0$成立，则SSR/σ^2与SSE/σ^2是独立的随机变量，且$SSR/\sigma^2\sim\chi^2(1), SSE/\sigma^2\sim\chi^2(n-2)$

使用检验统计量

$$F=\frac{SSR}{SSE/(n-2)}\sim F(1,n-2)$$

(1) $F_{0.01}(1,n-2)<F$，线性关系极其显著；
(2) $F_{0.05}(1,n-2)<F<F_{0.01}(1,n-2)$，线性关系显著；
(3) $F<F_{0.05}(1,n-2)$，无线性关系

3. Python实现

```
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm


def showErrorBar(result, datas):
    """绘制数据预测残差分布图"""
    pre = result.get_prediction(datas)
    df = pre.summary_frame(alpha=0.05)
    dfv = df.values
    low, upp = dfv[:, 4:].T  # 置信下限上限
    r = (upp-low)/2  # 置信半径
    num = np.arange(1, len(x0)+1)
    plt.errorbar(num, result.resid, r, fmt='o')


x0 = np.array([0.10, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.20, 0.22,
0.24])
y0 = np.array([42.0, 42.5, 45.0, 45.5, 45.0, 47.5, 49.0, 51.0, 50.0, 55.0, 57.5,
59.5])
datas = {'x': x0, 'y': y0}
result = sm.formula.ols('y~x', datas).fit()  # 拟合线性回归模型
print(result.summary())
print(result.outlier_test())  # 输出已知数据的野值检测
print('残差的方差:', result.mse_resid)  # 残差的方差
print(result.predict({"x": 0.2}))  # 预测
showErrorBar(result, datas)
plt.show()
```

```
>>> [OUT1]
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.977
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                     429.2
Date:                Mon, 07 Feb 2022   Prob (F-statistic):           1.52e-09
Time:                        16:53:02   Log-Likelihood:                -14.758
No. Observations:                  12   AIC:                             33.52
Df Residuals:                      10   BIC:                             34.49
```

```
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     28.4835      1.030     27.648      0.000      26.188      30.779
x            129.0094      6.227     20.716      0.000     115.134     142.885
==============================================================================
Omnibus:                        3.342   Durbin-Watson:                   2.356
Prob(Omnibus):                  0.188   Jarque-Bera (JB):                1.843
Skew:                          -0.957   Prob(JB):                        0.398
Kurtosis:                       2.840   Cond. No.                         24.4
==============================================================================
```

从上面打印的信息中，可以知道：

- 模型的一些基本信息【第一栏】

  观测数据（输入数据）的数量`No. Observations`，SSE的自由度`Df Residuals`，SSR的自由度`Df Model`

  判定系数R^2`R-squared`，显著性检验的F值`F-statistic`，显著性检验原假设H_0的概率`Prob (F-statistic)`

- 回归系数的信息【第二栏】

  系数的估计值`coef`，变量的显著性检验t值`t`，变量t检验原假设的概率`P>|t|`，置信区间`[0.025 0.975]`

```
>>> [OUT2]
    student_resid   unadj_p   bonf(p)
0        0.769358  0.461394  1.000000
1       -0.204792  0.842291  1.000000
2        1.284882  0.230915  1.000000
3        0.275629  0.789057  1.000000
4       -2.073621  0.067964  0.815569
5       -0.369701  0.720155  1.000000
6       -0.136740  0.894246  1.000000
7        0.655967  0.528257  1.000000
8       -2.418562  0.038702  0.464421
9        0.847083  0.418904  1.000000
10       0.794567  0.447310  1.000000
11       0.072405  0.943863  1.000000
```

上面打印的信息为数据的野值检测（最后一列），值越小表示该数据距数据整体的偏差越大，舍弃这些数据可以使拟合效果变好

```
>>> [OUT3]
残差的方差: 0.8221698113207546
```

上面打印的信息为残差的方差

```
>>> [OUT4]
0    54.285377
dtype: float64
```

上面打印的信息为预测值

二、多元线性回归模型

1. 模型建立与求解

>> 一般形式

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

>> 最小二乘估计

记

$$X=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1m}\\ 1 & x_{21} & x_{22} & \cdots & x_{2m}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm}\end{bmatrix},\quad Y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix},$$
$$\varepsilon=\begin{bmatrix}\varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n\end{bmatrix}^T,\quad \beta=\begin{bmatrix}\beta_0 & \beta_1 & \cdots & \beta_m\end{bmatrix}^T$$

- 线性回归方程：

$$Y=X\beta+\varepsilon$$

- 正规方程组

$$X^TX\beta=X^TY$$

- 参数估计值

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

- y的估计值

$$\hat{y}=b_0+b_1x_1+\cdots+b_mx_m$$

2. 模型分析

>> 一些参数的定义

- 原始数据y_i的变异程度

$$s_1^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$$

- 拟合值$\hat{y}_i$的变异程度

$$s_2^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2$$

- 正交分解式

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$
$$\Rightarrow\qquad SST=SSR+SSE$$

- **总变异平方和**：$SST=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=L_{yy}$

自由度：$df_T = n - m$

- **可解释的变异平方和**：$SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$

  自由度：$df_R = m$

- **残差平方和**：$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$

  自由度：$df_E = n - m - 1$

>> 相关性检验

【同一元线性回归】

>> 整体回归变量的显著性检验

原假设H_0：$\beta_1 = \beta_2 = \cdots = \beta_m = 0$

若假设H_0成立，则SSR/σ^2与SSE/σ^2是独立的随机变量，且
$SSR/\sigma^2 \sim \chi^2(m), SSE/\sigma^2 \sim \chi^2(n-m-1)$

使用检验统计量

$$F = \frac{SSR/m}{SSE/(n-m-1)} \sim F(m, n-m-1)$$

(1) $F_{0.01}(m, n-m-1) < F$，线性关系极其显著；
(2) $F_{0.05}(m, n-m-1) < F < F_{0.01}(m, n-m-1)$，线性关系显著；
(3) $F < F_{0.05}(m, n-m-1)$，无线性关系

>> 单个回归变量的显著性检验

原假设H_0：$\beta_j = 0,\ j = 0, 1, 2, \cdots, m$

若假设H_0：$\beta_j = 0$成立，使用检验统计量

$$t_j = \frac{b_j/\sqrt{c_{jj}}}{\sqrt{SSE/(n-m-1)}} \sim t(n-m-1)$$

其中，c_{jj}是$\left(X^T X\right)^{-1}$中的第(j, j)元素

(1) $|t_j| > t_{\frac{\alpha}{2}}(n-m-1),\ (j = 1, 2, \cdots, m)$，则$x_j$的作用显著；
(2) $|t_j| < t_{\frac{\alpha}{2}}(n-m-1),\ (j = 1, 2, \cdots, m)$，则$x_j$的作用不显著

三、主成分多元线性回归模型

1. 模型概述

主成分回归分析是为了克服最小二乘(LS)估计在数据矩阵A存在多重共线性时表现出的不稳定性而提出的。

主成分回归分析采用的方法是将原来的回归自变量变换到另一组变量，即主成分，选择其中一部分重要的主成分作为新的自变量，丢弃了一部分影响不大的主成分，实际上达到了降维的目的，然后用最小二乘法对选取主成分后的模型参数进行估计，最后再变换回原来的模型求出参数的估计。

2. 算法步骤

(1) 自变量标准化处理；
(2) 使用PCA降维，得到新的一组自变量；
(3) 对新的自变量做回归分析，得到回归系数；
(4) 将回归系数还原回标准化处理前的原自变量下的系数

3. 回归系数还原公式及其推导

>> 符号说明

- 原自变量X【$m \times n$】，主成分自变量Z【$m \times t$】，因量Y【$1 \times m$】
- 原自变量的均值向量μ【$1 \times n$】、方差向量σ【$1 \times n$】
- 主成分系数矩阵Σ【$t \times n$】
- 原自变量下的回归系数A【$1 \times n$】、b【常数】
- 主成分自变量下的回归系数A'【$1 \times t$】、b'【常数】

>> 还原公式

$$
\begin{aligned}
A &= \frac{A'\Sigma}{\sigma} \\
b &= b' - A\mu^T
\end{aligned}
$$

>> 还原公式推导

$$
\begin{aligned}
Y &= b' + A'Z^T \\
&= b' + A'\Sigma\left(\frac{X^T - \mu^T}{\sigma}\right) \\
&= b' - \frac{A'\Sigma}{\sigma}\mu^T + \frac{A'\Sigma}{\sigma}X^T \\
&= b + AX^T
\end{aligned}
$$