第六章 回归分析

实验8 回归分析及R语言实现（１）

目 录

8.1 实验目的

- 1. 熟练掌握使用R语言建立多元线性回归的方法；
- 2. 熟练掌握使用R语言建立逐步回归方程的方法。

8.2 实验原理

1、线性回归模型及模型参数β, σ^2的最小二乘估计

设Y是一可观测的随机变量，它受到$p-1$个非随机因素$X_1, X_2, \cdots, X_{p-1}$和随机误差$\varepsilon$的影响。假定它们有如下线性关系：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +, \cdots, +\beta_{p-1} X_{p-1} + \varepsilon$$

其中：β_k是待估参数，$\varepsilon \sim N(0, \sigma^2)$，则称上式为线性回归模型。

设对总体$(X_1, X_2, \cdots, X_{p-1}, Y)$进行$n$次（$n \geq p$）独立观测，得样本：

$$(x_{i,1}, x_{i,2}, \cdots, x_{i,p-1}, y_i) \quad i = 1,2, \cdots, n$$

令$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}$，$Y = (y_1 \quad \cdots \quad y_n)^T$，$\beta = (\beta_0 \quad \cdots \quad \beta_{p-1})^T$

则的最小二乘估计：$S(\hat{\beta}) = \min \varepsilon^T \varepsilon = \min (Y - X\beta)^T (Y - X\beta)$

可得正规方程：$X^T X \vec{\beta} = X^T Y$

若$rank(X)=p$，则有$\hat{\beta}=(X^TX)^{-1}X^TY$，代入线性回归方程，并略去误差项

的经验回归方程：$\hat{Y}=\widehat{\beta}_0+\widehat{\beta}_1X_1+\widehat{\beta}_2X_2+,\cdots,+\widehat{\beta}_{p-1}X_{p-1}$

令$\vec{e}=\vec{y}-\hat{y}$称为残差向量，可得：$E(\vec{e}^T\vec{e})=\sigma^2(n-p)$，$\hat{\sigma}^2=\frac{1}{n-p}\vec{e}^T\vec{e}$

则$\hat{\sigma}$是σ^2的无偏估计。

2、回归模型检验原理

（1）线性回归关系的显著性检验

为检验Y与$X_1, X_2, \cdots, X_{p-1}$之间是否存在显著的线性回归关系，即检验假设：

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

$$H_1: \text{至少有某个} \beta_i \neq 0 (1 \leq i \leq p-1)$$

构造如下检验统计量：$F = \frac{MSR}{MSE}$

当H_0为真时，$F \sim F(p-1, n-p)$；给定显著性水平α，由F分布得临界值$F_\alpha(p-1, n-p)$(即F分布的上侧α分位数)，计算F的观测值F_0，若$F_0 \leq F_\alpha(p-1, n-p)$，接受$H_0$，否则拒绝$H_0$。对显著性检验问题，其输出结果通常是检验的$p$值。对上述线性回归关系的显著性检验问题，检验的$p$值$P_{H_0}\{F \geq F_0\}$，若$p$值小于显著性水平$\alpha$，拒绝$H_0$，否则接受$H_0$。

（2）回归参数的显著性检验

回归关系显著并不意味着每个自变量X_i对Y的影响都显著，我们希望从回归方程中剔除那些对Y的影响不显著的的自变量，从而建立一个较为简单有效的回归方程。若某个自变量X_k对Y无影响，那么它的系数$\beta_k = 0$，因此检验X_k的影响是否显著等价与检验假设：$H_0: \beta_k = 0 \leftrightarrow H_1: \beta_k \neq 0$。若令$S(\hat{\beta}) = MSE(X^T X)^{-1}$，$s(\hat{\beta}_K)$为$S(\hat{\beta})$的主对角线上的第$k$个元素的平方根，则可得到：

$$\frac{\widehat{\beta_k} - \beta_k}{s(\widehat{\beta_k})} \sim t(n-p)$$

当H_0为真时，

$$t = \frac{\widehat{\beta_k}}{s(\widehat{\beta_k})} \sim t(n-p)$$

记t的观测值为t_0，则检验准则为：

$$\begin{cases} \text{若}|t_0| \leq t_{\frac{\alpha}{2}}(n-p), \text{则接受}H_0 \\ \text{若}|t_0| > t_{\frac{\alpha}{2}}(n-p), \text{则拒绝}H_0 \end{cases}$$

（3）残差分析

在回归分析中，我们通常假定$\varepsilon_i(i=1,2,\cdots,n)$是独立同正态分布的随机变量，有零均值和常值方差$\sigma^2$，因此，若拟合的回归模型适合于所给的数据，那么残差$\varepsilon_i(i=1,2,\cdots,n)$应该基本上反映误差的这些特性。利用残差的这些特性反过来考察原模型的合理性就是残差分析的基本思想。

① 残差正态性的频率检验

回归模型中标准化残差

$$\frac{e_i}{\sqrt{MSE}} \quad (i=1,2,\cdots,n)$$

可近似认为是取自标准正态总体的样本，理论上属于68%在(-1,1)内，87%在(-1.5,1.5)内，95%在(-1,1)内，如果残差在某个区间内的频率与上述理论频率有较大的偏差，我们有理由怀疑e_i(从而ε_i)的正态假设的合理性。

② 残差正态性的QQ图检验

QQ图是做正态性检验的直观方法，将残差$\varepsilon_i(i = 1,2,\cdots,n)$按由小到大的排列，以残差为纵坐标、正态期望为横直角坐标系中画出正态QQ图。

③ 相关系数检验法

通过计算残差和正态期望之间的相关系数判断它们之间关系的强弱，若相关系数接近1，则说明残差为正态性

④ 时序残差图分析

- 以观测时间(或观测值序号)为横坐标，的散点图时序残差图。拟合好的模型的时序残差图中的点应落在以时间轴为中轴线的带状区域，且无明显的趋势性，否则说明回归方程的形式或对误差等方差的存在一定问题。
- 以拟合值$\hat{Y}$为横坐标的残差图分析

若模型适当，以拟合值$\hat{Y}$为横坐标的残差图

- 以自变量为横坐标的残差图分析
- 以每个X_j的各观测值x_{ij}为点的横坐标，以残差为纵坐标。同样满意的残差图呈现水平带状。

3、逐步回归的原理与步骤

逐步回归的基本步骤就是依次拟合一系列回归方程，后一个回归方程在前一个的基础上增加或删除一个自变量，其增加或删除某个自变量的准则是用残差平方和的增加或减少量来衡量，一般采用如下的偏F检验统计量，设模型已经有$l-1$个自变量，记 个自变量的集合为A，当不再A中的自变量加入到模型当中时，偏F检验统计量一般形式为：

$$F=\frac{SSE(A)-SSE(S,X_k)}{\frac{SSE(A,X_k)}{n-l-1}}=\frac{SSR(X_k|A)}{MSE(A,X_k)}\sim F(1,n-l-1)$$

$SSR(X_k|A)=SSE(A)-SSE(S,X_k)$称为额外回归平方和。$F$统计量描述了误差平方和的增加或减少量，所以偏$F$检验统计量是逐步回归方法中增加或删除变量所用的基本统计量。

8.3 实验过程

1、多元线性回归

在R中，可以通过函数lm()来求解多元回归问题。lm()用法如下：

lm(formula,data,subset,weights,na.action,method = "qr",model = TRUE,

x = FALSE,y = FALSE,qr = TRUE,singular.ok = TRUE,contrasts = NULL,offset,...)

- formula ：公式，形如y~x；
- data ：数据框，由样本数据构成；
- subset ：可选项，表示所选用的样本子集；
- weights ：可选向量，表示对应样本的权重；

- na.action ：函数，表示当数据中出现缺失数据时的处理方法；
- singular.ok ：逻辑变量，取FALSE表示奇异值拟合是错误的；