第四章 回归分析

4.1 一元线性回归分析

回归分析 回归分析是用来寻找若干变量之间**统计联系**（关系）的一种方法。它是一种统计模型，分为**线性回归**和**非线性回归**，线性回归在气象中最为常用（解释性好，物理机理较为清晰）。
利用回归分析得到的统计关系对某一变量作出未来时刻的估计，称为**预报值(量)**。**前期**（也可以是同期因子）已发生的多个与之有关的气象要素称为**预报因子**。

案例分析

为了预报某地某月平均气温或降水量情况（预报量），选择预报前期已发生的多个有关的气象要素（预报因子），利用回归分析方法分析多个预报因子和预报变量之间的相互关系，建立统计关系的方程式，最后利用其对未来时刻的气温或降水量作出预报估计。

一元回归 一元回归分析处理的是**两个变量**之间的关系，即一个预报量和一个预报因子之间的关系。

4.1.1 回归模型

基本原理 对抽取容量为 n 的预报量 y 与预报因子 x 的一组样本（必须保证样本个数一致），**若认为 y 与 x 是一元线性统计关系**，则线性回归方程为：$y_i = b_0 + bx_i + \varepsilon_i$，$i = 1,2,\cdots,n$（$\varepsilon_i$ 为残差项，我们希望它越小越好），那么预报量的估计量 $\hat{y}$ 与 x 有如下关系：

$$\hat{y}_i = b_0 + bx_i \quad i = 1,2,\cdots,n$$

或写为一般的回归方程：$\hat{y} = b_0 + bx$，其中b_0为截距，b为斜率。

最小二乘法 对所有的 x_i，**若 $\hat{y}_i$ 与 y_i 的偏差最小**，就认为所确定的直线能最好地代表所有实测点的散布规律。
为了消除偏差符号的影响，可以用**偏差的平方**来反映偏差的绝对值偏离情况。
全部观测值与回归直线的**离差平方和**记为：

$$Q(b_0,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

它刻画了全部观测值与回归直线的偏离程度。显然Q值越小越好，Q是待定系数b_0和b的函数。

标准方程组 根据**极值原理**，要求：$\frac{\partial Q}{\partial b_0} = 0$，$\frac{\partial Q}{\partial b} = 0$。整理得到求回归系数$b_0$、$b$的方程组：

$$\begin{cases} nb_0 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

称为求回归系数的**标准方程组**。

具体求解

① $\frac{\partial Q}{\partial b_0} = \frac{\partial}{\partial b_0}\left(\sum_{i=1}^{n}(y_i - b_0 - bx_i)^2\right) = \sum_{i=1}^{n} -2(y_i - b_0 - bx_i) = 0 \Rightarrow \sum_{i=1}^{n}(-y_i - b_0 - bx_i) = 0 \Rightarrow$

$nb_0 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ 其中 $\frac{\partial(y_i - b_0 + bx_i)^2}{\partial b_0} = 2(y_i - b_0 + bx_i)\cdot(-1)$。

② $\frac{\partial Q}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=1}^{n}(y_i - b_0 - bx_i)^2\right) = -2\sum_{i=1}^{n} x_i(y_i - b_0 - bx_i) = 0 \Rightarrow b_0\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$

回归系数 $\boxed{b_0=\bar{y}-b\bar{x}}$ $\boxed{b=\frac{\sum_{i=1}^n x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2-n\bar{x}^2}=\frac{S_{xy}}{S_x^2}}$

距平形式 将 $b_0=\bar{y}-b\bar{x}$ 代入回归方程 $\hat{y}_i=b_0+bx_i$，得到 $\hat{y}_i-\bar{y}=b(x_i-\bar{x})$ 或 $\hat{y}_{di}=bx_{di}$

标准化形式 发现有关系：$b=\frac{S_{xy}}{S_x^2}=\frac{S_{xy}}{S_xS_y}\cdot\frac{S_y}{S_x}=r_{xy}\cdot\frac{S_y}{S_x}$，由此 $\hat{y}_{zi}=r_{xy}x_{zi}$ （这里的x,y都是标准化后的变量）

相关系数 回归系数b与相关系数之间的关系：$b=\frac{S_{xy}}{S_x^2}=r_{xy}\cdot\frac{S_y}{S_x}$ 相关系数r与回归系数b同号

当$b<0$，回归直线斜率为负，预报量 y 随预报因子 x 增加而减少，反映预报量与因子是负相关。
当$b>0$，回归直线斜率为正，预报量 y 随预报因子 x 增加而增加，反映预报量与因子是正相关。

4.1.2 回归问题的方差分析

意义 评价回归方程的优劣

预报量方差 **预报量方差**可以表示成**回归估计值的方差**（回归方差）和**误差**（残差）方差之和：

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

即 $s_y^2=s_{\hat{y}}^2+s_e^2$。

评估分析 方差分析表明，预报量 y 的变化可以看成由前期因子 x 的变化所引起的，同时加上随机因素 e 变化的影响，这种前期因子 x 的变化影响可以用回归方差的大小来衡量。如果**回归方差大/残差方差小**，表明用线性关系解释y与x的关系比较符合实际情况，**回归模型比较好**。

离差平方和 $\frac{1}{n}\sum_{i=1}^n(y_i-\bar{y})^2=\frac{1}{n}\sum_{i=1}^n(\hat{y}_i-\bar{y})^2+\frac{1}{n}\sum_{i=1}^n(y_i-\hat{y}_i)^2$ 两边同时乘以n变成各变量离差平方和的关系。

总离差平方和：$s_{yy}=U+Q=\sum_{i=1}^n(y_i-\bar{y})^2$ 反映**因变量 y 的 n 个观测值与其均值的总离差**
回归平方和：$U=\sum_{i=1}^n(\hat{y}_i-\bar{y})^2$ 反映**回归值的分散程度**
残差平方和：$Q=\sum_{i=1}^n(y_i-\hat{y})^2$ 反映**观测值偏离回归直线的程度**

4.1.3 相关系数与线性回归

判决系数 因为回归方差不可能大于预报量的方差，可以用它们的比值来衡量方程的拟合效果。即：

$$\frac{s_{\hat{y}}^2}{s_y^2}=\frac{\frac{1}{n}\sum_{i=1}^n(\hat{y}_i-\bar{y})^2}{\frac{1}{n}\sum_{i=1}^n(y_i-\bar{y})^2}=\frac{U}{s_{yy}}=r_{xy}^2$$

上式表明预报因子x对预报量y的方差的线性关系程度，这一比值又称为**回归方程判决系数/解释方差**。
判决系数是衡量两个变量线性关系密切程度的量，等于两变量相关系数的平方。
如果是多元线性回归，合理猜想，是复相关系数的平方。

物理含义
① **回归平方和占总离差平方和的比例** ② 反映回归直线的**拟合程度**
③ 取值范围在[0,1] ④ 判决系数**等于相关系数的平方**
⑤ $r^2\to1$ 说明回归方程拟合的越好，$r^2\to0$ 说明回归方程拟合的越差

4.1.4 回归方程的显著性检验

中心思想 显著性检验的主要思想是检验预报因子与预报量**是否有线性关系**。

统计量 可以证明在**原假设总体回归系数为 0** 的条件下，统计量：

$$F=\frac{U/1}{Q/(n-2)}$$

遵从**分子自由度为1，分母自由度为$(n-2)$**的F分布。

显著性检验 查 F 的分布表，在 $\alpha=0.05$ 下，若 $F>F_\alpha$ 则认为回归方程是显著的。反之，则不显著。

相关系数 统计量F也可以写为：$F=\frac{U/1}{Q/(n-2)}=\frac{s_{\hat{y}}^2/1}{s_e^2/(n-2)}=\frac{r^2}{(1-r^2)/(n-2)}$，与 $t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ 比较，发现二者等价。

注意 F的相关系数表达式开方就是相关系数t检验的表达式，故**一元回归方程的检验与相关系数的检验一致**。

4.1.5 回归系数的显著性检验

说明　气象中使用最多的是回归方程的**距平形式**，所以对回归方程的显著性检验可以**只对因子的回归系数**进行检验。

统计量　在**原假设H_0: 回归系数$\beta=0$**的条件下，统计量 $t=\dfrac{b-\beta}{\sqrt{\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}=\dfrac{b/\sqrt{c}}{\sqrt{Q/(n-2)}}$ 遵从自由度为 $n-2$ 的 t 分布.

其中：$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{Q}{n-2}$　　$c=\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{-1}$

或者根据F分布与t分布的关系，统计量 $F=\dfrac{U/1}{Q/(n-2)}=\dfrac{b^2/c}{Q/(n-2)}$　遵从分子自由度为1，分母自由度为 $n-2$的 F 分布。其中 $U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(b_0-bx_i-b_0-b\bar{x})^2=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2=\dfrac{b^2}{c}$

4.1.6 预报值的置信区间

置信区间　因为 $e_i=y_i-E(y_i)$ 可以看成遵从 **$N(0,\sigma^2)$ 的正态分布**，所以其 **95% 的置信区间为 $E(y_i)\pm1.96\sigma$**

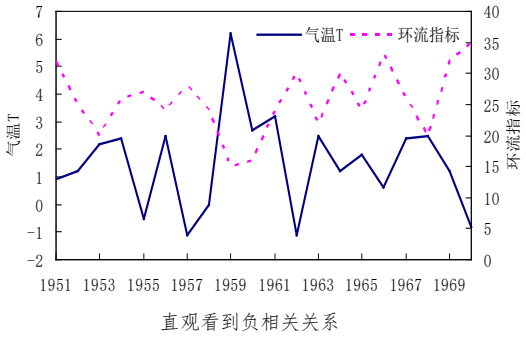
$E(y_i)$ 可用 $\boldsymbol{b_0+bx_i=\hat{y}_i}$ 估计，σ 可用无偏估计量 $\hat{\sigma}=\sqrt{\dfrac{Q}{n-2}}$ 估计，$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

预报值的 95% 置信区间可近似估计为 $[\hat{y}_i-1.96\hat{\sigma},\hat{y}_i+1.96\hat{\sigma}]$。

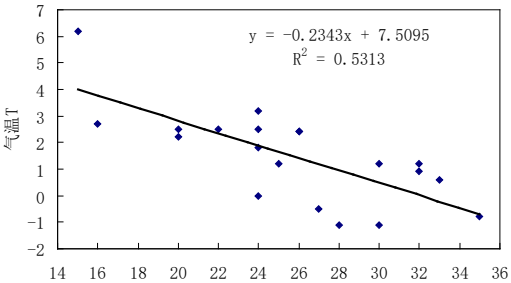
每一个点的置信区间都不一样，置信区间上下界是一个曲线。

4.1.7 一元线性回归分析预测步骤

分析数据



直观看到负相关关系



由环流指标推算气温

第一步　**计算回归系数，确定方程**。对上述资料，容易算得 $n=20$，$\sum_{i=1}^{\infty}x_i=513$，$\sum_{i=1}^{\infty}y_i=30.0$，$\sum_{i=1}^{1}x_i^2=13721$，$\sum_{i=1}^{2}x_iy_i=637$　根据 $b_0=\bar{y}-b\bar{x}$，$b=\dfrac{\sum_{i=1}^{n}x_iy_i-n\overline{xy}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=\dfrac{S_{xy}}{S_x^2}$　可以解得：

$b_0=7.5$，$b=-0.23$　最终得到回归方程：$\hat{y}=7.5-0.23x$

第二步　**回归方程显著性检验**。

再次计算得到：$\sum_{i=1}^{1}y_i^2=103.12$ 于是 $r_{xy}=\sqrt{\dfrac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\sqrt{\dfrac{13721-\frac{1}{20}\times(513)^2}{103.12-\frac{1}{20}\times(30.0)^2}}\times(-0.23)=-0.727$

最终得到：$F=\dfrac{(-0.727)^2}{[1-(-0.727)^2]/(20-2)}=20.18$ 查询F分布表，在$\alpha=0.05$，分子自由度为 1，分母自由度为 18 时，$F_\alpha=4.41$ 由于 $F>F_\alpha$ 认为回归方程是显著的。（考试可以灵活应用t检验）

第三步　**计算预报值的置信区间，作出预测**。

将$x=24$代入回归方程，计算出预报值为 $y_{24}=1.98$℃，又有 $Q=s_{yy}-U=s_{yy}-s_{yy}r^2=s_{yy}(1-r^2)$

算出：$\hat{\sigma}=\sqrt{\frac{1}{20-2}\times58.12(1-0.727^2)}=1.11$ ，用 $E(y_i)\pm1.96\sigma$ 得到置信区间。

所以 1971 年北京 3 月下旬气温的 95%置信区间为−0.2~4℃。