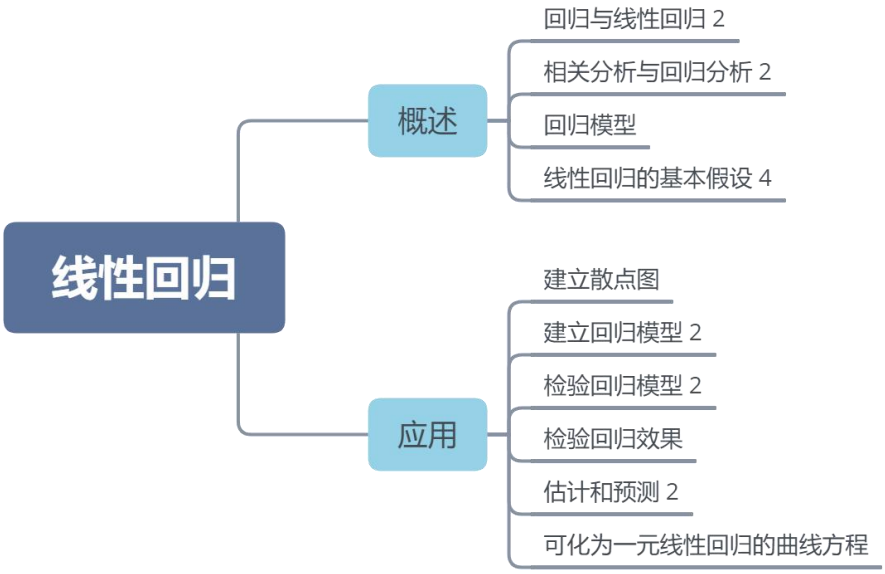
第十二章　线性回归



一、概述

（一）回归与线性回归

1.回归分析

通过大量的观测发现变量之间存在的统计规律性，并用一定的数学模型表示变量相关关系的方法。

2.一元线性回归分析

当只有一个自变量，并且统计量大体是一次函数的线性关系的回归分析。

（二）相关分析和回归分析

区别：

1.相关分析

（1）用相关系数来度量变量间的密切程度；

（2）相关分析是双向的，不强调哪个是自变量，哪个是因变量。

2.回归分析

（1）旨在用数学模型来表示变量之间数量关系的可能形式；

（2）回归分析是单向的，要找出一个变量随着另一个或多个变量的变化而变化的关系。

联系：

从广义上而言，相关分析包括回归分析，两者的共同点是确定变量之间是否存在关系，另外在一元线性回归中，相关系数等于两回归系数的几何平均数：

$$b_{YX} = r \times \frac{S_Y}{S_X},\ b_{XY} = r \times \frac{S_X}{S_Y}$$

（三）回归模型

用来表达变量之间规律的数学模型即为回归模型。

一元线性回归用$\hat{Y}$=a+bX 作为回归方程，代表 X 与 Y 的线性关系，其中：

X：自变量，Y：对应于 X 的 Y 变量的估计值；

a：表示该直线在 Y 轴的截距；

b：表示该直线的斜率，即 X 变化时 Y 的变化率，表示 X 变化一个单位时，$\hat{Y}$变化 b 个单位。又叫作 Y 对 X 的回归系数，用 by.x 或 X→Y 表示。

若以 Y 做自变量，回归方程变为：$\hat{X}$=a+bY，这时

b：Y 变化时 X 的变化率，表示 Y 变化一个单位，$\hat{X}$变化 6 个单位。又叫 X 对 Y 的回归系数，用 bx.y 或 Y→X 表示。

（四）线性回归的基本假设

设回归方程为：$\hat{Y}$=a+bX

1.线性关系假设

x 与 Y 在总体上具有线性关系，这是线性回归的最基本假设。

2.正态性假设

在回归分析中的 Y 服从正态分布。

3.独立性假设

（1）一个 X 对应的 Y 值与另一个 X 对应的 Y 值间独立。

（2）不同 X 产生的误差相互独立，误差与 X 间也相互独立。

4.误差等分散性假设

X 对应的误差，除呈随机化的常态分布，其变异量也应相等，称为误差等分散性。

二、回归模型的综合应用

（一）建立散点图

根据直接搜集的数据资料做出散点图，直观判断两变量间是否大致呈直线关系。

（二）建立回归模型

1.平均数法

设$\hat{Y}$=a+bX，将数据按照奇偶分成两组，然后分别代入回归方程，形成二元一次方程组后分别解出 a 和 b。

2.最小二乘法

（1）原理

散点图中每一点沿 Y 轴方向到直线的距离（Y-$\hat{Y}$）的平方和最小，即误差的平方和最小。

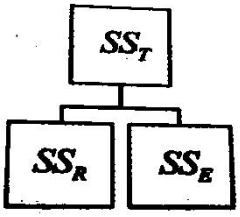
（2）计算

$$a=\overline{Y}-b\overline{X},\quad b=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}\qquad b_{YX}=r\times\frac{s_Y}{s_X},\quad b_{XY}=r\times\frac{s_X}{s_Y}$$

（三）检验回归模型

1.方差分析

利用方差分析的方法对回归模型进行有效性检验。



（1）平方和

$$SS_T = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

$$SS_R = \sum (\hat{Y} - \bar{Y})^2 = b^2 (\sum X^2 - \frac{(\sum X)^2}{N})$$

$$SS_E = SS_T - SS_R$$

（2）自由度

$df_T=df_R+df_E$，$df_T=N-1$，$df_E=N-2$，$df_R=df_T-df_E=1$

（3）F 值

$$F = \frac{MS_R}{MS_E}$$

（4）字母含义

SS_T：所有 Y 值的总平方和，SS_R：由回归直线表示的线性关系导致的变异，SS_E：误差变异。

2.回归系数进行显著性检验

利用假设检验的方法对回归系数进行显著性检验，样本回归系数服从 t 分布，使用 t 检验：

$$SE_b = \sqrt{\frac{s_{YX}^2}{\sum (X - \bar{X})^2}} \qquad s_{YX}^2 = \frac{\sum (Y - \hat{Y})^2}{N - 2} = MS_E$$

$$t = \frac{b - \beta}{SE_b}, \quad df = n - 1$$

其中，SE_b为回归系数的标准误
Syx为误差的标准误

（四）检验回归效果

采用决定系数（r^2）来衡量回归效果：r^2 等于回归平方和在总平方和中所占的比例，该比重越大，误差平方和在总离差平方和中占的分量就越小。在回归分析中，自变量所决定的

离差平方和（回归平方和）在总离差中所占的比例越大越好。因此，可以把测定系数作为回归有效性高低的指标。若 $r^2=0.64$，则表明 Y 的变异中有 64%是由变量 X 的变异引起的，或说有 64%可由 X 的变异解释。

决定系数的公式：SS_R/SS_T

（五）估计和预测

1.点预测

直接将确定的自变量 X_i 的值带人回归模型，得到相应的 Y 值。

2.区间预测

以一定的概率为保证，预测当自变量 X 取定的值 X_i 时， 因变量 Y 的可能范围。

$$\hat{Y}-t_{\alpha/2}s_{YX}<Y<\hat{Y}+t_{\alpha/2}s_{YX}$$

$$s_{YX}^2=\frac{\sum(Y-\hat{Y})^2}{N-2}=MS_E$$

（六）可化为一元线性回归的曲线方程

当变量之间的关系不是线性的，而是非线性（曲线）的关系时，一个基本思路就是设法将非线性关系线性化，然后用线性回归模型进行处理。

可化为一元线性回归的曲线模型主要有多项式模型、指数模型、幂函数模型、对数模型和成长曲线模型。