第八讲（第12章）回归分析预测

Regression and Forecasting

内容提要

- 函数关系、相关关系的基本概念
- 相关分析与回归分析的基本概念
- 相关分析
- 一元线性回归分析与预测
- 多元线性回归分析与预测

函数关系与相关关系

- 函数关系是指事物间的数量变化关系可以用函数关系式表示的确定性关系。即自变量的每一个确定的X值，因变量总有一个唯一的确定的Y值与之对应。
- 相关关系，亦称非确定性关系。它是指变量之间相互关系中不存在数值对应关系的非确定性的依存关系。

相关关系的种类

- 按相关的程度可分为完全相关、不完全相关和不相关。一般的相关现象是不完全相关。
- 按相关的方向可分为正相关和负相关。
- 按相关的形式可分为线性相关和非线性相关。
- 按变量多少可分为单相关、复相关和偏相关。一个变量对另一变量的相关关系，称为单相关。一个变量对两个以上变量的相关关系时，称为复相关。在某一现象与多种现象相关的场合，当假定其他变量不变时，其中两个变量的相关关系称为偏相关。

相关分析与回归分析

- 相关分析是用一个指标来表明现象间依存关系的密切程度。
- 回归分析是用数学模型近似表达变量间的平均变化关系。
- 相关分析可以不必确定变量中哪个是自变量，哪个是因变量，其所涉及的变量都是随机变量。
- 回归分析必须事先确定具有相关关系的变量中哪个为自变量，哪个为因变量。一般地说，回归分析中因变量是随机的，而把自变量作为研究时给定的非随机变量。
- 一定要始终注意把定性分析和定量分析结合起来，在定性分析的基础上开展定量分析。

相关分析与回归分析的用途

- 相关分析用于研究两个变量之间联系的强度。如销售额与广告支出的关系，消费者对质量的认知是否与其对价格的认知有关。
- 回归分析被广泛用于解释市场份额、销售额、品牌偏好的差异，以及用广告、价格、分销和产品质量等营销管理变量解释其他的营销结果。

相关分析

- 一、单相关系数及其检验
- 二、复相关系数和偏相关系数

一、单相关系数及其检验

- 总体相关系数定义：

$$\rho = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

- 总体相关系数反映两变量之间线性相关程度的一种特征值。

一、单相关系数及其检验

- （一）样本相关系数（Pearson相关系数）的定义

$$r = \frac{\sum (X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum (X_t - \bar{X})^2 \sum (Y_t - \bar{Y})^2}}$$

- 样本相关系数的定义还可从另一个角度给出。设Y倚X和X倚Y的样本回归方程为：

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t$$

$$\hat{X}_t = \hat{\alpha}_1 + \hat{\alpha}_2 Y_t$$

样本相关系数可定义为样本回归系数的乘积的开方，即:

$$r = \pm \sqrt{\hat{\beta}_2 \hat{\alpha}_2}$$

上式中 r 的符号应与回归系数的符号一致。

- （二）相关系数与可决系数
- 简单线性回归模型中相关系数 r 的平方等于可决系数 r^2。
-

- （三）单相关系数的检验
- H_0: ρ =0
- 统计量 $$r\sqrt{\frac{n-2}{1-r^2}}$$
- 在零假设下服从自由度为n-2的t分布。

二、复相关系数和偏相关系数

- （一）复相关系数

$$R = \frac{\sum (Y_t - \bar{Y})(\hat{Y}_t - \bar{Y})}{\sqrt{\sum (Y_t - \bar{Y})^2 \sum (\hat{Y}_t - \bar{Y})^2}}$$

实际计算复相关系数时，一般是先计算出可决系数，然后再求可决系数的平方根。复相关系数只取正值。

- （二）偏相关系数

计算偏相关系数时，需要掌握多个变量的数据，一方面考虑多个变量之间可能产生的影响，一方面又用一定的方法控制其他变量，专门考察两个特定变量的净相关关系。偏相关系数与单相关系数数值上可能相差很大，甚至符号都可能相反。

Partial correlation

- 偏相关：两变量在排除其他变量影响之后的相关。

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

- $r_{12.3}$是变量1和2在移除变量3对她们的影响之后的相关
- 比较$r_{12.3}$和r_{12}两者间是否有差异
  - 无显著差异存在时表示变量3对变量1和2无影响
- 如果$r_{12.3}=0$但是$r_{12}>0$
  - 变量1和2的相关是假相关
- 通常用来做因果的推论

一元线性回归分析与预测

- 一、标准的一元线性回归模型
- 二、一元线性回归模型的估计
- 三、一元线性回归模型的检验
- 四 、一元线性回归模型预测

“回归”一词的历史渊源

- 回归一词最先由加尔顿（Francis Galton)引入。在一篇论文中，加尔顿发现，虽然有一个趋势，父母高，儿女也高；父母矮，儿女也矮，但给定父母的身高，儿女辈的平均身高却趋向于或者“回归”到全体人口的平均身高。
- 换言之尽管父母双亲都异常高或异常矮，而儿女的身高则有走向人口总体平均身高的趋势。

“回归”一词的历史渊源

- 加尔顿的普遍回归定律（law of universal regression)还被他的朋友皮尔逊（Pearson)证实。皮尔逊曾收集过一些家庭群体的1千多名成员的身高记录。他发现，对于一个父亲高的群体，儿辈的平均身高低于他们父辈的身高。这样就把高的和矮的儿辈一同“回归”到所有男子的平均身高。用加尔顿的话说，这是“回归到中等（regression to mediocrity)”。

回归的现代释义

- 回归分析是关于研究一个叫做应变量（被解释变量）的变量对另一个或多个叫做解释变量（自变量）的变量的依赖关系，其用意在于通过后者（在重复抽样中）的已知或设定值，去估计和（或）预测前者的（总体）均值。

一元线性回归的步骤

绘制散点图Plot the Scatter Diagram

↓

拟定一般模型Formulate the General Model

↓

估计参数Estimate the Parameters

↓

估计标准化回归系数 Standardized Regression Coefficients

↓

系数显著性检验Test for Significance

↓

对方程整体显著性进行拟合优度检验

↓

检查预测精度Check Prediction Accuracy

↓

检查残差Examine the Residuals

↓

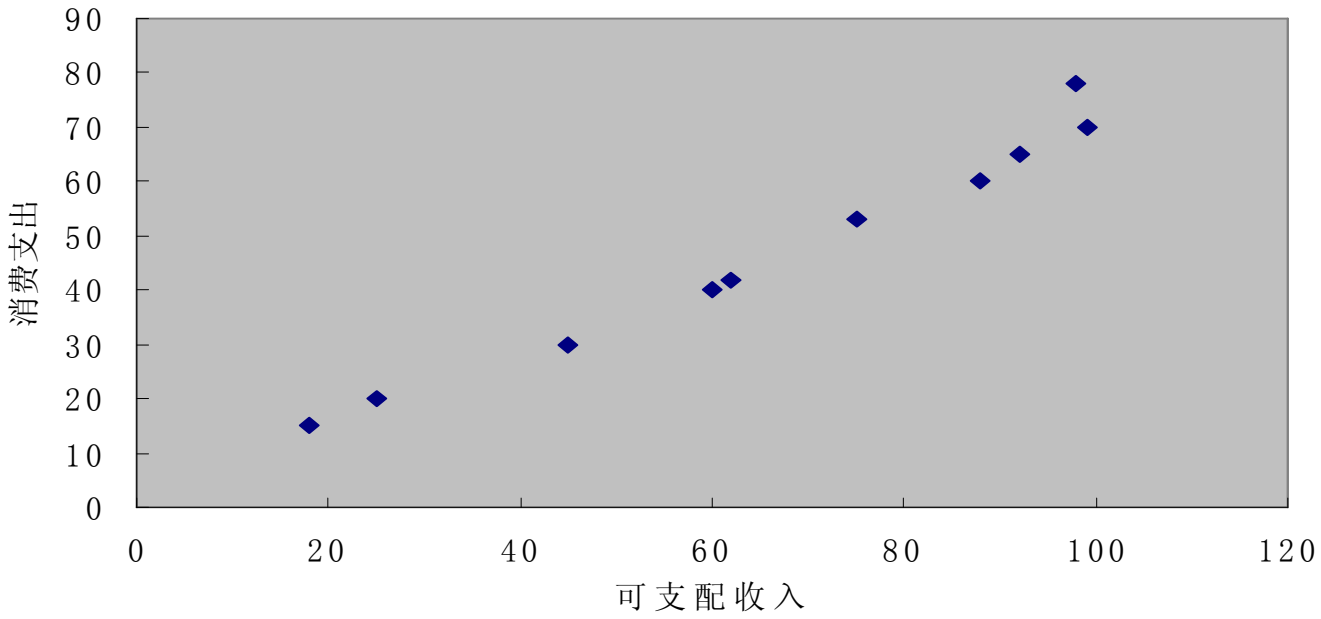
交叉验证Cross-Validate the Model

居民消费和收入的相关表

C	15	20	30	40	42	53	60	65	70	78
I	18	25	45	60	62	75	88	92	99	98

散点图（scattergram)

图9-1 消费与收入的相关图



一、标准的一元线性回归模型

- (一)总体回归函数

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

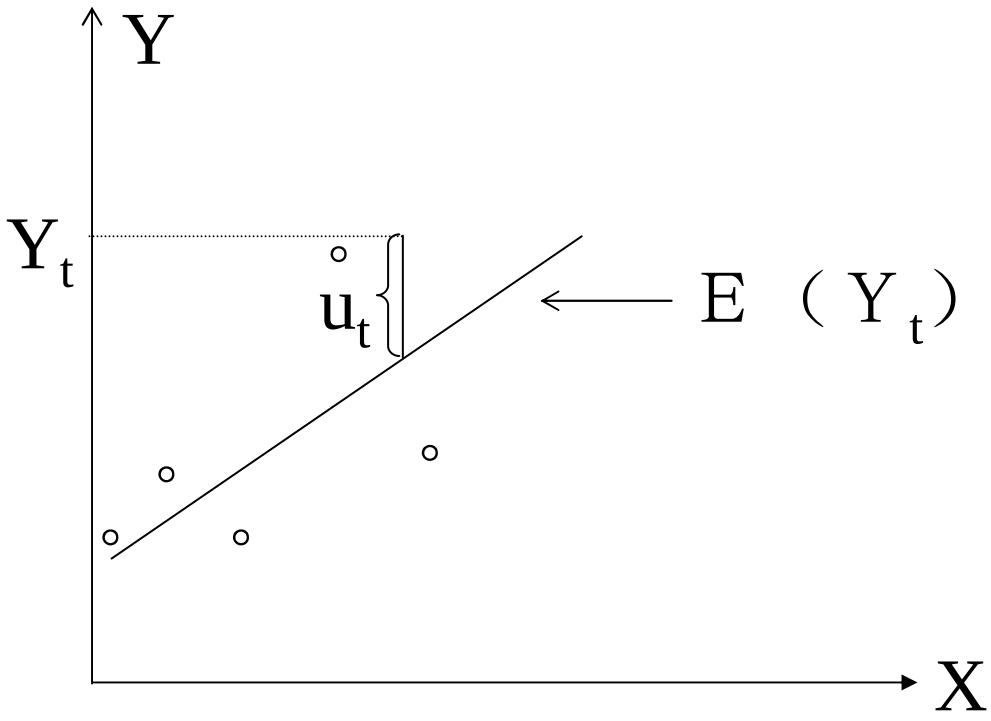
u_t是随机误差项，又称随机干扰项，它是一个特殊的随机变量，反映未列入方程式的其他各种因素对Y的影响。

- (二)样本回归函数:

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X + e_t \quad t = 1, 2, \ldots n$$

e_t称为残差，在概念上，e_t与总体误差项u_t相互对应；n是样本的容量。

总体回归线与随机误差项



$$E(Y_t) = \beta_1 + \beta_2 X_t$$

随机干扰项的意义

- 1、理论的含糊性。即使有决定Y的行为的理论，也是不完全的。
- 2、数据的欠缺。例如，要找解释家庭消费支出的变量，一般得不到关于家庭财富的信息。
- 3、核心变量与随机变量。
- 4、人类行为的内在随机性。
- 5、节省原则。

样本回归函数与总体回归函数区别

- 总体回归线是未知的，只有一条。样本回归线是根据样本数据拟合的，每抽取一组样本，便可以拟合一条样本回归线。
- 总体回归函数中的 β_1和 β_2是未知的参数，表现为常数。而样本回归函数中的$\hat{\beta}_1$和$\hat{\beta}_{2t}$ 是随机变量，其具体数值随所抽取的样本观测值不同而变动。
- 总体回归函数中的u_t是 Y_t与未知的总体回归线之间的纵向距离，它是不可直接观测的。而样本回归函数中的 e_t是 Y_t与样本回归线之间的纵向距离，当根据样本观测值拟合出样本回归线之后，可以计算出 e_t的具体数值。

误差项的标准假定

- 假定 1： $E(u_t) = 0$
- 假定 2： $Var(u_t) = E(u_t^2) = \sigma^2$
- 假定 3： $Cov(u_t u_s) = E(u_t u_s) = 0 \quad t \neq s$
- 假定 4：自变量是给定变量，与误差项线性无关。
- 假定 5：随机误差项服从正态分布。
- 满足以上标准假定的一元线性回归模型，称为标准的一元线性回归模型。

二、一元线性回归模型的估计

- （一）回归系数的估计　最小二乘法

设 $Q=\sum e_t^2=\sum(Y_t-\hat{Y}_t)^2=\sum(Y_t-\hat{\beta}_1-\hat{\beta}_2X_t)^2$

将Q对求偏导数，并令其等于零，可得:

$$-2\sum(Y_t-\hat{\beta}_1-\hat{\beta}_2X_t)=0$$

$$-2\sum X_t(Y_t-\hat{\beta}_1-\hat{\beta}_2X_t)=0$$

加以整理后有：

- $$n\hat{\beta}_1+\hat{\beta}_2\sum X_t=\sum Y_t$$

$$\hat{\beta}_1\sum X_t+\hat{\beta}_2\sum X_t^2=\sum X_tY_t$$

回归系数的最小二乘估计量

- 以上方程组称为正规方程组或标准方程组，式中的 n 是样本容量。
- 求解这一方程组可得：

$$\hat{\beta}_2 = \frac{n\sum X_t Y_t - \sum X_t Y_t}{n\sum X_t^2 - (\sum X_t)^2}$$

$$\hat{\beta}_1 = \sum Y_t / n - \hat{\beta}_2 \sum X_t / n = \bar{Y} - \hat{\beta}_2 \bar{X}$$

三、一元线性回归模型的检验

- （一）方程整体显著性的拟合程度评价
- 总离差平方和的分解

ＳＳＴ＝ＳＳＲ＋ＳＳＥ

ＳＳＴ是总离差平方和；ＳＳＲ是回归平方和；ＳＳＥ是残差平方和。

- 可决系数：

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

- 可决系数的特性

（二）系数显著性检验

- 1. 提出假设。
- 2.确定显著水平α 。
- 3.计算回归系数的 t 值。

$$t_{\hat{\beta}_2} = \frac{\hat{\beta}_2 - \beta_2^*}{S_{\hat{\beta}_i}}$$

- 4.确定临界值。
- 双侧检验查 t 分布表所确定的临界值是（$-t_{\alpha}/2$）和（$t_{\alpha}/2$）；单侧检验所确定的临界值是（t_{α}）。
- 5.做出判断。

四、一元线性回归模型预测

- （一） 简单回归预测的基本公式：

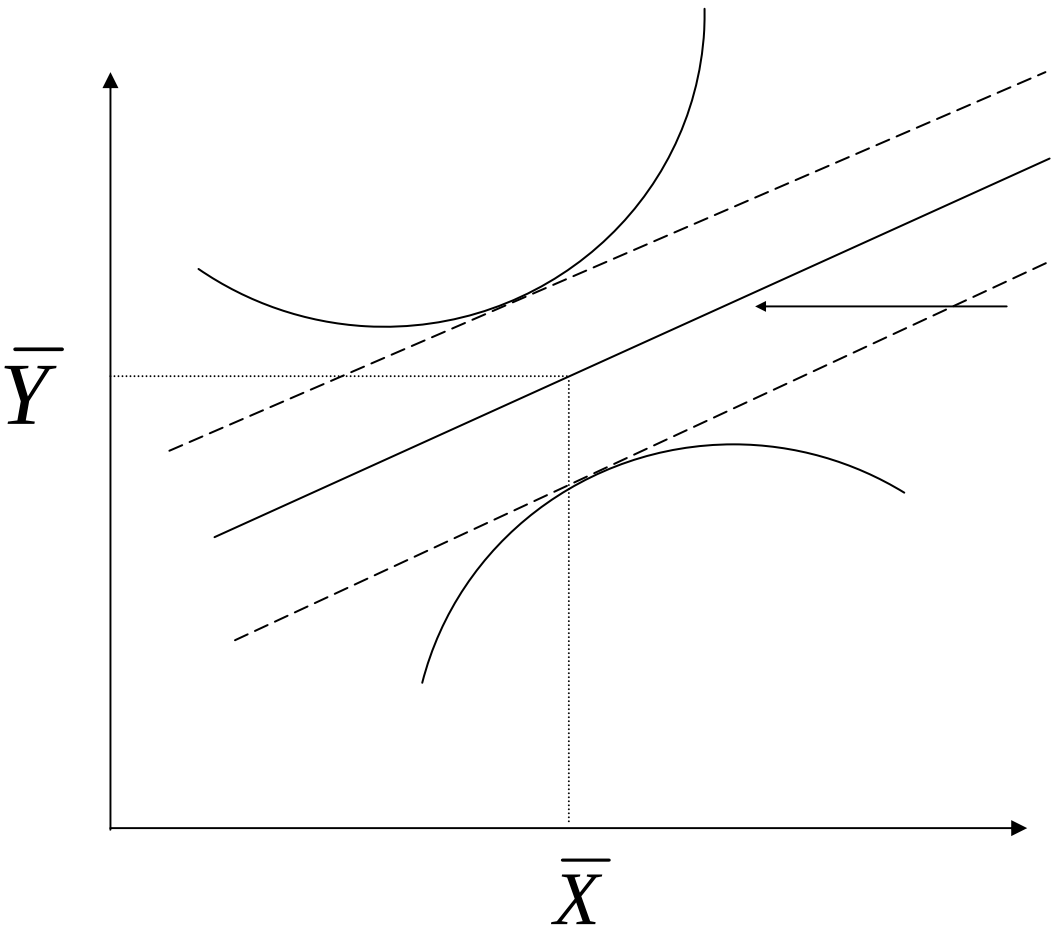
$$\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f$$

- 回归预测是一种有条件的预测，在进行回归预测时，必须先给出 X_f的具体数值。内插检验或事后预测。外推预测或事前预测。
- （二）预测误差
- 发生预测误差的原因。
- 预测误差 $Var(e_f) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum (X_t - \bar{X})^2}\right)$
- （三）区间预测

  Y_f的（1－α）的置信区间为：$Y_f \pm t_{\alpha/2}(n-2) \times Se_f$
- 回归预测的置信区间的特点。

$\bar{Y}$

回归预测的置信区间



$$\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f$$

多元线性回归分析与预测

- 一、标准的多元线性回归模型
- 二、多元线性回归模型的估计
- 三、多元线性回归模型的检验和预测
- 四、多元线性回归预测

一、标准的多元线性回归模型

- 多元线性回归模型总体回归函数的一般形式

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t$$

- 多元线性回归模型的样本回归函数

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \ldots + \hat{\beta}_k X_{kt} + e_t$$

多元线性回归分析的标准假定除了包括上一节中已经提出的的假定外，还要追加一条假定。这就是回归模型所包含的自变量之间不能具有较强的线性关系。

二、多元线性回归模型的估计

- （一）回归系数的估计
- $\hat{B} = (X'X)^{-1}X'Y$
- （二）总体方差的估计
- $S^2 = \dfrac{\sum e_t^2}{n-k}$
- （三）最小二乘估计量的性质
- 标准的多元线性回归模型中，高斯.马尔可夫定理同样成立。

三、多元线性回归模型的检验和预测

- （一）拟合程度的评价
- 修正自由度的可决系数（理由）。

$$= 1 - \frac{\sum e_t^2 /(n-k)}{\sum (Y_t - \bar{Y})^2 /(n-1)}$$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-k)}$$

式中，n 是样本容量；k 是模型中回归系数的个数。

- 修正自由度的可决系数 $\bar{R}^2$ 的特点。

(二)显著性检验

- 1. 回归系数的显著性检验

$$t = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \quad j=1, 2, \cdots, k$$

式中， $S_{\hat{\beta}_j}$ 是的标准差的估计值。 按下式计算：

- $$S_{\hat{\beta}_j} = \sqrt{S^2 \times \psi_{jj}}$$

- 式中，ψ_{jj} 是$(X'X)^{-1}$的第 j 个对角线元素， S^2是随机误差项方差的估计值。式的 t 统计量的原假设是H_0： $\beta_j = 0$ ，因此 t 的绝对值越大表明β_j为 0 的可能性越小，即表明相应的自变量对因变量的影响是显著的。

2．回归方程的显著性检验

- 具体的方法步骤
- 回归模型方差分析表

离差名称	平　方　和	自由度	方　差
回归平方和	SSR= $\sum(\hat{Y}_t-\bar{Y})^2$	k−1	SSR/(k−1)
残差平方和	SSE= $\sum e_t^2$	n−k	SSE/(n−k)
总离差平方和	SST= $\sum(Y_t-\bar{Y})^2$		

- （3）F 统计量

$$F=\frac{SSR/(k-1)}{SSE/(n-k)}$$

四、多元线性回归预测

- 基本公式：
- $$\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_{2f} + \ldots + \hat{\beta}_k X_{kf}$$

式中，X_{jf}(j=2,3,……k)是给定的X_j在预测期的具体数值；$\hat{\beta}_j$ 是已估计出的样本回归系数；$\hat{Y}_f$是X_j给定时 Y 的预测值。