第6章 相关分析与回归分析

§ 6.1 相关分析

§ 6.2 简单线性回归分析

§ 6.3 多元回归分析应用

§6.1 相关分析

- 关系形态：

  确定性函数关系：假设有两个变量X 和Y，当变量X取某个数值时，变量Y按某种法则有唯一确定的一个数值与之对应，则称Y为X的函数。

  统计相关关系 既有一定关系但又不唯一确定的关系。

  例如高收入家庭的消费一般来说要比低收入的家庭高一些，但同样人均月收入8000元的两个家庭，他们的人均消费可能差异很大。

- 关系方向：正相关、负相关
- 相关形态：曲线相关、直线相关

简单相关系数

样本相关系数 $r_{XY}=\dfrac{\sum X_iY_i-\frac{1}{n}\sum X_i\sum Y_i}{\sqrt{\sum X_i^2-\frac{1}{n}(\sum X_i)^2}\times\sqrt{\sum Y_i^2-\frac{1}{n}(\sum Y_i)^2}}$

相关程度

完全相关: $|r|=1$ 充分必要条件为

$$Y_i=a+bX_i \qquad i=1,2,\cdots,n$$

高度相关: $|r|>0.8$

显著相关: $|r|$ 在 0.5~0.8 之间

低度相关: $|r|$ 在 0.3~0.5 之间

不 相 关: $|r|<0.3$

注：样本相关系数衡量的是两指标间的线性相关程度

案例分析

- 设总体表示某地死于癌症人数X(万人)和钢铁产量Y(万吨)，近5年内的观测值见教材。

$$r = \frac{22.6 - 2 \times 2.8}{\sqrt{14.8 - 2^2}\ \sqrt{35.6 - 2.8^2}} = 0.9819$$

- 这两个指标，从数量上看高度相关，但显然，死于癌症人数和钢铁产量高度相关的结论是不合理的。

注　意

- （1）注意独立与不相关的区别，因此 $r=0$， 只能说明X和Y没有线性关系，但不能说明没有其它（函数）关系，以下不特殊说明，相关指的线性相关。
- （2）若X和Y从事物的性质看，几乎没有相互关系，但其相关系数却很大，这种伪相关，人们很容易识别与避免，但2个有关系的经济指标的相关系数的大小，是否一定就反映两者的相关程度，这种程度上的“伪”，却容易忽略。

- （3）能使用简单相关系数衡量2个指标线性相关程度的条件是“当被考虑的2个变量不受其余变量的影响，或其余变量对这2个变量的综合影响几乎为零时，可用简单相关系数衡量2个变量的线性相关程度”。

  因此，在经济数量分析中，简单相关系数仅有有限的参考价值。

偏相关

- 在研究由个*k*个相互作用影响、制约的变量 $X_1,\cdots,X_k$ 决定的经济系统中，衡量Xi 和Xj 之间的线性关系，要用偏相关系数衡量，它反映的是固定其余k-2个变量后或扣除其余k-2个变量对Xi和Xj的影响后，Xi和Xj之间的关系。

偏相关2

- 当只有X_i, X_j和X_k 3个变量时，X_i和X_j的偏相关系数为

$$r_{ij\bullet k} = \frac{r_{ij} - r_{ik} r_{jk}}{\sqrt{(1-r_{ik}^2)(1-r_{jk}^2)}}$$

- 其中r_{ij}等是相应变量的简单相关系数，表达式无不反映扣除X_i和X_k，X_j和X_k的相关影响。

案例分析（续）

- 前面我们看到死于癌症人数X和钢铁产量Y样本的简单相关系数很大，但不一定可以说他们之间存在很强的线性关系，下面我们将添加时间t作为新变量，求X和Y在扣除时间*t*的影响后的偏相关系数.

$r_{xt}=0.990, r_{yt}=0.993, r_{xy}=0.982$

则扣除时间*t*的影响后， X和Y的的偏相关系数为

$$r_{XY\bullet t}=\frac{0.990-0.990*0.993}{\sqrt{(1-0.99\text{^}2)(1-0.993\text{^}2)}}=-0.072$$

复相关系数

- 设是维(y,X)是k+1维总体，其中X是维随机向量，通常称y与X的任意线性组合的相关系数的最大值为y与X的复相关系数。即

$$r^* = \max_{a} \rho_{Y,aX}$$

因此，复相关系数是衡量1个变量与1组变量之间线性相关程度的量。

用R表示样本复相关系数作为r*的估计，R的平方就是后面讲到的关于多自变量的回归模型中的R^2 。

案例：消费、收入间的相关

	上海消费	上海收入	全国消费	全国收入
上海消费	0.999	0.999	0.999	0.998
上海收入	0.839	0.998	0.996	0.997
全国消费	0.673	−0.640	0.999	0.999
全国收入	−0.385	0.588	0.896	0.999

表中对角线元素是复相关系数的平方R_j^2的值；表的下三角是偏相关系数；上三角是相关系数。例如上海收入关于上海消费、全国消费和全国收入的复相关系数的平方是0.998；上海收入与全国消费的相关系数是0.996，偏相关系数是-0.641

典型相关

在经济数量分析中，有时需要度量2组变量间的相互关系，如1组内生变量与1组外生变量间的关系，又如1组先行指标与1组一致指标间的关系，这时可用典型相关系数来衡量。典型相关系数定义为2组变量各自任意线性组合的最大相关系数，并且相应的线性组合的系数向量，可以用来构成综合指数的权重。

§6.2 简单线性回归分析

设有因变量Y与一组自变量 $X_1, X_2, \cdots, X_k$ 之间存在相关关系，这种关系可以表示为

$$Y = f(X_1, X_2, \cdots, X_k) + u$$

回归方程： $Y = f(X_1, X_2, \cdots, X_k)$

线性回归方程： $f(X_1, X_2, \cdots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

理论线性回归模型： $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u$

经验回归方程： $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$

- u反映:那些没有考虑进来的因素；

  观测误差等偶然因素对因变量及自变量的影响；非线性因素

随机项的一些假设

- 假设一：误差项是一个期望值为零的随机变量，即$Eu=0$。
- 假设二：在所有的时点t，u_t的方差都相同。
- 假设三：u_t是一个服从正态分布的随机变量，且相互独立。
- 注：假设一、二，称为马尔科夫假定，假设三称为正态假定。

线性回归模型的三种用途

- 1．结构分析：所谓结构分析，主要用于经济数量分析，包含两重意思，即研究分析经济变量之间的内在联系和检验经济理论。
- 2．预测：预测就是根据客观事物的过去和现在的发展规律，借助于科学的方法和技术手段，对未来的发展趋势和状况进行描述分析，形成科学的假设和判断。
- 3．政策模拟：政策模拟是经济数量分析中的提法，它有重要的应用，指一个决策者从众多决策中通过比较，选择一种最优政策来执行之，这一过程就是政策模拟，也是经济实验中的一种。

简单线性模型参数的最小二乘估计

对简单线性回归模型可以写为

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1,2,\cdots,n$$

给定的 β_0, β_1 模型残差平方和

$$L(\beta_0, \beta_1) = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

对上式求偏导数经整理可以得到正规方程组

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i = \sum Y_i$$

$$\hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 = \sum Y_i X_i$$

求解得到：

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum X_i Y_i - \sum X_i \sum Y_i / n}{\sum X_i^2 - (\sum X_i)^2 / n}, \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

案例分析

某厂全年利润与年初计划全年广告投入有着直接关系，过去6年的数据如下：

x	1.5	2	3	4	5	5.5	21
y	3	5	7	8	10	12	45
xy	4.5	10	21	32	50	66	183.5
x^2	2.25	4	9	16	25	30.25	86.5

列表求对的经验回归方程；并讨论该模型主要用于什么目的？

$$b=\frac{\sum xy-n\bar{x}\bar{y}}{\sum x^2-n\bar{x}^2}=\frac{183.5-6\times3.5\times7.5}{86.5-6\times3.5^2}=\frac{156}{78}=2$$

$$a=\bar{y}-\bar{x}b=7.5-3.5\times2=0.5 \quad \therefore \hat{y}=0.5+2x$$

广告投入是年初的计划，全年利润是年末的因此，模型可用于预测目的。

最小二乘估计的性质

- **性质1：** $\hat{\beta}_0, \hat{\beta}_1$ 都是的Y*i* 的线性函数，而且是参数的无偏估计。
- **性质2：** 在所有回归系数的线性无偏估计中，最小二乘估计具有最小方差 。
- **性质3：** $\hat{\sigma}^2 = \frac{1}{n-k-1} S_e^2$ 是随机项方差的无偏估计。若记
- $S_X^2 = \sum (X_i - \bar{X})^2$， $\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_X^2}$ ，$\mathrm{var}(\hat{\beta}_0) = \mathrm{var}(\hat{\beta}_1) \frac{1}{n} \sum X_i^2$

平方和分解

- Y总的离差平方和 $S_y^2 = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - (\sum Y_i)^2 / n$
- 回归平方和 $S_v^2 = \sum (\hat{Y}_i - \bar{Y})^2$

  是由自变量的变化所引起的变化，或者说是可以由回归直线来解释的的一部分变化
- 残差平方和 $S_e^2 = \sum (Y_i - \hat{Y}_i)^2$

  是不能用回归直线(自变量)来解释的部分

平方和分解

总离差平方和 = 回归平方和＋残差平方和

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

判决系数

- **定义：**回归平方和占总离差平方和的比例称为判决系数，记为R^2。

计算公式为 $$R^2=\frac{S_v^2}{S_y^2}=\frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}=1-\frac{S_e^2}{S_y^2}$$

当k=1时，残差平方和可利用已知结果计算

$$S_e^2=\sum y^2-a\sum y-b\sum xy$$

判决系数一定在0~1之间，越接近1说明回归直线模拟样本数据越好，也可说自变量解释因变量的能力越强。

模型总体效果检验

- 原假设：$H_0:\beta_1=\cdots=\beta_k=0$，　$H_1:\beta_j$ 不全为零

  检验统计量为

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

当原假设成立时有 $F\sim F(k,n-k-1)$

如果$F>F_a(k,n-k-1)$拒绝原假设，说明Y与k个自变量X的线性关系显著。对于上例，由于$k=1,F=69.43>7.71=F_{0.05}(1,4)$，模型效果好，利润与广告投入有很强的（线性）关系。

参数的显著性检验

- 通过了总体效果的检验，只能说明Y与k个自变量X从整体上看线性关系显著，并不表明每个X_j都与Y有显著的线性关系，因此还需检验每个X_j是否显著，换句话说，就是X_j的系数b_j是否显著不为零，作进一步检验。
- k=1时问题是要检验原假设

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0 \qquad j = 0,1$$

- 其检验统计量是

$$t_j = \frac{\hat{\beta}_j}{V(\hat{\beta}_j)} \sim t(n-k-1)$$

其中$V(b_j)$是b_j标准差

如果t_j的绝对值>$t_{a/2}(n-k-1)$则拒绝原假设，认为X_j作用显著，解释Y的能力强。

预测

- 回归分析的目的之一是根据所建立的回归方程进行预测与控制。在回归方程通过各种检验之后，就可以利用经验回归方程来实现预测与控制的目标。
- 1、点预测

如果我们得到了经验回归方程 $Y = \hat{\beta}_0 + \hat{\beta}_1 X$

已知自变量的特定值X_0，利用经验回归方程可求得Y_0的点估计值。例如，在例6.4中，如果下一年度预算投入6千万元的广告费用，则该年企业利润的预测值为

$Y_0 = 0.5 + 2 \times 6 = 12.6$（千万元）。

- 2、区间预测（略）

案例分析：消费模型

- 在研究上海人均消费水平的问题中，记因变量Y为上海人均消费金额(元)，记自变量X为人均可支配收入(元)，样本数据1985~2004年20年，其数据见教材表6.2
- 人均消费与收入的散点图 近似在一条直线上，可建立两者的线性模型。
- 在宏观经济学中消费的理论模型是

$$Y=b_0+b_1X+u$$

- 其中b_0是自主消费倾向，表示收入为零时人均所需（最低）消费的金额，b_1是边际消费倾向，反映收入增加1元将增加的多少元用于消费。

消费模型(续)

- 经计算可以得到如下结果：

$$Y=376.85+0.7267X$$

（4.963）　　（82.134）

$$R^2=0.997,\ F=5982$$

- 模型各种检验值表明，这20年来上海人均消费与收入间存在极强的长期均衡关系，上海平均自主消费为376.85元(收入为零时的必须消费，也可作为贫困线的参考值)，边际消费倾向是0.7267，即收入每增加1元，平均增加0.7267元用于消费。

案例分析：宏观经济政策

- 我国1992~2003年国内生产总值GDP，货币供应量M1、国家财政支出ZC和社会消费品零售总额XF的数据见教材表6.5，希望根据这些数据建立一个GDP关于M1、ZC和XF的回归方程，并用它来分析财政政策、货币政策及消费需求对国内生产总值的影响。
- 为消除偶然因素影响，先对每组数据作3项移动平均处理，并补充首尾数据。对处理后数据，通过计算得到如下的回归方程：

宏观经济政策续(1)

- Y=1633.75-0.12M1+1.79ZC+7.34XF

（0.40）　（-1.34）　（0.59）　（12.80）

R^2=0.999，F=2089.003

从整体上讲，所选的几个变量基本上能够解释GDP的变化。除了社会消费品零售总额的系数肯定通过检验之外，其余几项都不能通过系数的显著性检验。基于我们的研究目的以及常数项的值最小，先删除常数项(注：不是同时删除所有不显著变量)，重新计算回归方程得到：

宏观经济政策续(2)

- Y=-0.158M1+2.941ZC+7.666XF

  （-7.335）　（3.425）　（95.604）

  R^2=0.999，F=15298

- 基于模型的建立，还是可以得到不少的结论。
- 1．新模型相对原模型有极大改进。
- 2．我国GDP与财政政策、货币政策和消费需求之间存在极强的内在联系(均衡关系)，我国宏观经济可以通过各项政策进行调控。
- 3．对我国GDP增长影响最大、最显著的是消费需求，并且，财政政策比货币政策对我国的经济影响要大。
- 4．模型所反映的是综合影响，不能说货币对GDP是负相关。