第10章　回归分析

介绍：

1、回归分析的概念和模型

2、回归分析的过程

回归分析的概念

- 寻求有关联（相关）的变量之间的关系
- 主要内容：
  - 从一组样本数据出发，确定这些变量间的定量关系式
  - 对这些关系式的可信度进行各种统计检验
  - 从影响某一变量的诸多变量中，判断哪些变量的影响显著，哪些不显著
  - 利用求得的关系式进行预测和控制

回归分析的模型

- 按是否线性分：线性回归模型和非线性回归模型
- 按自变量个数分：简单的一元回归，多元回归
- 基本的步骤：利用**SPSS**得到模型关系式，是否是我们所要的，要看回归方程的显著性检验（**F**检验）和回归系数**b**的显著性检验**(T**检验**)**，还要看拟合程度R^2**(**相关系数的平方**,**一元回归用**R Square**，多元回归用**Adjusted R Square)**

回归分析的过程

- 在回归过程中包括：
  - **Liner：线性回归**
  - **Curve Estimation：曲线估计**
  - **Binary Logistic： 二分变量逻辑回归**
  - **Multinomial Logistic：多分变量逻辑回归**
  - **Ordinal 序回归**
  - **Probit：概率单位回归**
  - **Nonlinear：非线性回归**
  - **Weight Estimation：加权估计**
  - **2-Stage Least squares：二段最小平方法**
  - **Optimal Scaling 最优编码回归**
- 我们只讲前面3个简单的（一般教科书的讲法）

10.1 线性回归(Liner)

- 一元线性回归方程: $y=a+bx$
  - a称为截距
  - b为回归直线的斜率
  - 用R^2判定系数判定一个线性回归直线的拟合程度：用来说明用自变量解释因变量变异的程度（所占比例）
- 多元线性回归方程: $y=b_0+b_1x_1+b_2x_2+\ldots+b_nx_n$
  - b_0为常数项
  - b_1、b_2、…、b_n称为y对应于x_1、x_2、…、x_n的偏回归系数
  - 用Adjusted R^2调整判定系数判定一个多元线性回归方程的拟合程度：用来说明用自变量解释因变量变异的程度（所占比例）
- 一元线性回归模型的确定:一般先做散点图(Graphs ->Scatter->Simple),以便进行简单地观测（如：Salary与Salbegin的关系）
- 若散点图的趋势大概呈线性关系，可以建立线性方程，若不呈线性分布，可建立其它方程模型，并比较R^2 (-->1)来确定一种最佳方程式（曲线估计）
- 多元线性回归一般采用逐步回归方法-Stepwise

逐步回归方法的基本思想

- 对全部的自变量$x_1,x_2,\cdots,x_p$,按它们对Y贡献的大小进行比较，并通过F检验法，选择偏回归平方和显著的变量进入回归方程，每一步只引入一个变量，同时建立一个偏回归方程。当一个变量被引入后，对原已引入回归方程的变量，逐个检验他们的偏回归平方和。如果由于引入新的变量而使得已进入方程的变量变为不显著时，则及时从偏回归方程中剔除。在引入了两个自变量以后，便开始考虑是否有需要剔除的变量。只有当回归方程中的所有自变量对Y都有显著影响而不需要剔除时，在考虑从未选入方程的自变量中，挑选对Y有显著影响的新的变量进入方程。不论引入还是剔除一个变量都称为一步。不断重复这一过程，直至无法剔除已引入的变量，也无法再引入新的自变量时，逐步回归过程结束。

10.1.6 线性回归分析实例p240

- 实例：**P240Data07-03** 建立一个以初始工资**Salbegin** 、工作经验**prevexp** 、工作时间**jobtime** 、工作种类**jobcat** 、受教育年限**edcu**等为自变量，当前工资**Salary**为因变量的回归模型。
- 先做数据散点图,观测因变量**Salary**与自变量**Salbegin**之间关系是否有线性特点
- **Graphs ->Scatter->Simple**
- **X Axis： Salbegin**
- **Y Axis： Salary**
- 若散点图的趋势大概呈线性关系，可以建立线性回归模型
- **Analyze->Regression->Linear**
- **Dependent: Salary**
- **Independents: Salbegin,prevexp,jobtime,jobcat,edcu**等变量
- **Method: Stepwise**
- 比较有用的结果：
- 拟合程度**Adjusted R2**： 越接近**1**拟合程度越好
- 回归方程的显著性检验**Sig**

10.2 曲线估计(Curve Estimation)

- 对于一元回归，若散点图的趋势不呈线性分布，可以利用曲线估计方便地进行线性拟合(liner)、二次拟合(Quadratic)、三次拟合(Cubic)等。采用哪种拟合方式主要取决于各种拟合模型对数据的充分描述(看修正Adjusted R^2 -->1)

不同模型的表示		
模型名称	回归方程	相应的线性回归方程
Linear(线性)	$Y=b_0+b_1t$	
Quadratic(二次)	$Y=b_0+b_1t+b_2t^2$	
Compound(复合)	$Y=b_0(b_1^t)$	$Ln(Y)=ln(b_0)+ln(b_1)t$
Growth(生长)	$Y=e^{b0+b1t}$	$Ln(Y)=b_0+b_1t$
Logarithmic(对数)	$Y=b_0+b_1ln(t)$	
Cubic(三次)	$Y=b_0+b_1t+b_2t^2+b_3t^3$	
S	$Y=e^{b0+b1/t}$	$Ln(Y)=b_0+b_1/t$
Exponential(指数)	$Y=b_0*e^{b1*t}$	$Ln(Y)=ln(b_0)+b_1t$
Inverse(逆)	$Y=b_0+b_1/t$	
Power(幂)	$Y=b_0(t^{b1})$	$Ln(Y)=ln(b_0)+b_1ln(t)$
Logistic(逻辑)	$Y=1/(1/u+b_0b_1^t)$	$Ln(1/Y-1/u)=ln(b_0+ln(b_1)t)$

10.2.3 曲线估计(Curve Estimation)分析实例

- 实例**P247 Data11-01**：有关汽车数据，看**mpg**(每加仑汽油行驶里程)与**weight**(车重)的关系
  - 先做散点图(**Graphs ->Scatter->Simple**)：**weight(X)**、**mpg(Y)**，看每加仑汽油行驶里程数**mpg(Y)**随着汽车自重**weight(X)**的增加而减少的关系，也发现是曲线关系
  - 建立若干曲线模型（可试着选用所有模型**Models**)
    - **Analyze->Regression-> Curve Estimation**
    - **Dependent: mpg**
    - **Independent: weight**
    - **Models:** 全选(除了最后一个逻辑回归)
    - 选**Plot models**：输出模型图形
    - 比较有用的结果：各种模型的**Adjusted R^2**，并比较哪个大，结果是指数模型**Compound**的**Adjusted R^2=0.70678**最好（拟合情况可见图形窗口），结果方程为：$mpg=60.15*0.999664^{weight}$
    - 说明：**Growth**和**Exponential**的结果也相同，也一样。

10.3二项逻辑回归(Binary Logistic)

- 在现实中，经常需要判断一些事情是否将要发生，候选人是否会当选？为什么一些人易患冠心病？为什么一些人的生意会获得成功？此问题的特点是因变量只有两个值，不发生(0)和发生(1)。这就要求建立的模型必须因变量的取值范围在0～1之间。
- **Logistic回归模型**
  - **Logistic模型**：在逻辑回归中，可以直接预测观测量相对于某一事件的发生概率。包含一个自变量的回归模型和多个自变量的回归模型公式：

$$prob(event) = \frac{1}{1+e^{-z}}$$

其中： $z=B_0+B_1X_1+\ldots B_pX_p$(P为自变量个数）。某一事件不发生的概率为Prob(no event)＝1-Prob(event) 。因此最主要的是求B_0,B_1,…B_p(常数和系数)

  - 数据要求：因变量应具有二分特点。自变量可以是分类变量和定距变量。如果自变量是分类变量应为二分变量或被重新编码为指示变量。指示变量有两种编码方式。
  - 回归系数：几率和概率的区别。几率=发生的概率/不发生的概率。如从52张桥牌中抽出一张A的几率为(4/52)/(48/52)=1/12，而其概率值为4/52=1/13

根据回归系数表，可以写出回归模型公式中的z。然后根据回归模型公式 Prob(event) 进行预测。