概述

(一)相关关系

(1)函数关系:

如:销售额与销售量;圆面积和圆半径.

(2)统计关系:

如:收入和消费;身高的遗传.

概述

统计关系的常见类型：

- 线性相关：正线性相关、负线性相关
- 非线性相关

统计关系不象函数关系那样直接,但却普遍存在,且有强有弱.如何测度?

概述

(二) 相关分析和回归分析的任务

- 研究对象:统计关系
- 相关分析旨在测度变量间线性关系的强弱程度.
- 回归分析侧重考察变量之间的数量变化规律,并通过一定的数学表达式来描述这种关系,进而确定一个或几个变量的变化对另一个变量的影响程度.

相关分析

(一)目的

通过样本数据,研究两变量间线性相关程度的强弱.(例如:职工的年龄和收入之间的关系、工人数和管理人员之间的数量关系)

(二)基本方法

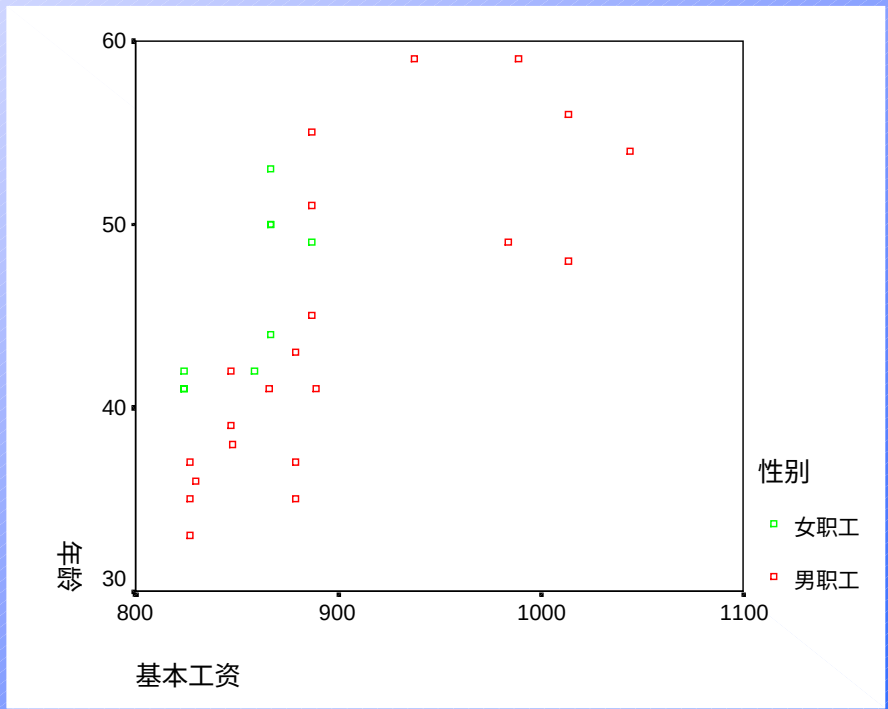
绘制散点图、计算相关系数

绘制散点图

(一)散点图

将数据以点的形式绘制在直角平面上.比较直观,可以用来发现变量间的关系和可能的趋势.

体现了正相关趋势

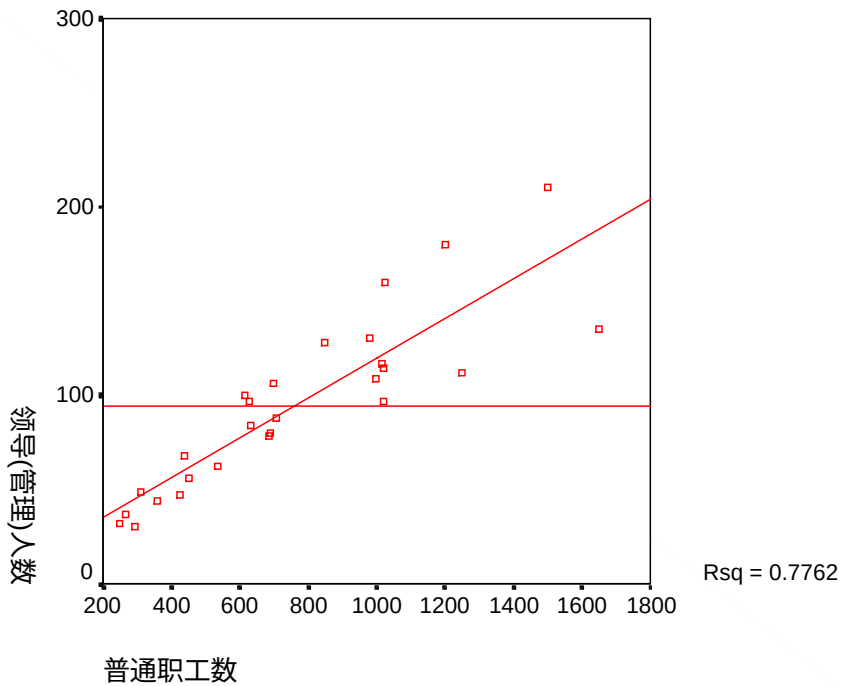


绘制散点图

(二)应用举例

通过27家企业普通员工人数和管理人员数,利用散点图分析人数之间的关系

散点图在进行相关分析时较为粗略



计算相关系数

(一)相关系数

(1)作用:

- 以精确的相关系数(r)体现两个变量间的线性关系程度.
- r:[-1,+1]; r=1:完全正相关; r=-1:完全负相关; r=0:无线性相关; |r|>0.8:强相关; |r|<0.3:弱相关

计算相关系数

(一)相关系数

(2)说明:

- 相关系数只是较好地度量了两变量间的线性相关程度,不能描述非线性关系.
- 数据中存在极端值时不好

计算相关系数

(一)相关系数

(3)种类:

- 简单线性相关系数(Pearson)

  (如:身高和体重)

- Spearman相关系数和Kendall相关系数:

(如:不同年龄段与不同收入段,职称和受教育年份)

(4)相关系数检验

计算相关系数

(二)应用举例

- 通过27家企业普通员工人数和管理人员数,利用相关系数分析人数之间的关系
  - *表示t检验值发生的概率小于等于0.05,即总体无相关的可能性小于0.05;
  - **表示t检验值发生的概率小于等于0.01,即总体无相关的可能性小于0.01;
  - **比*更严格.

偏相关分析

(一) 含义

在控制了其他变量的影响下计算两变量的相关系数。

(二)应用举例

分析商品的需求量(y)和该商品的价格(x1)及消费者的收入(x2)的关系

回归分析概述

(一)回归分析理解

(1)“回归”的含义
(2)回归的第一类含义
(3)回归的第二类含义

回归分析概述

(二)回归分析的基本步骤

(1)确定自变量和因变量

(2)从样本数据出发确定变量之间的数学关系式,并对回归方程的各个参数进行估计.

(3)对回归方程进行各种统计检验.

(4)利用回归方程进行预测.

线性回归分析概述

(三)参数估计的准则

- 目标:回归线上的预测值与观察值之间的距离总和达到最小
- 最小二乘法(利用最小二乘法拟和的回归直线与样本数据点在垂直方向上的偏离程度最低)

一元线性回归分析

(一)一元回归方程:

- $y=\beta_0+\beta_1 x$
- β_0为常数项；β_1为y对x回归系数，即:x每变动一个单位所引起的y的平均变动

(二)一元回归分析的步骤

- 利用样本数据建立回归方程
- 回归方程的拟和优度检验
- 回归方程的显著性检验(t检验和F检验)
- 残差分析
- 预测

一元线性回归方程的检验

(一)拟和优度检验

(1)目的:检验样本观察点聚集在回归直线周围的密集程度，评价回归方程对样本数据点的拟和程度。

(2)统计量：判定系数

(二)回归方程的显著性检验

(1)目的:检验自变量与因变量之间的线性关系是否显著，是否可用线性模型来表示.

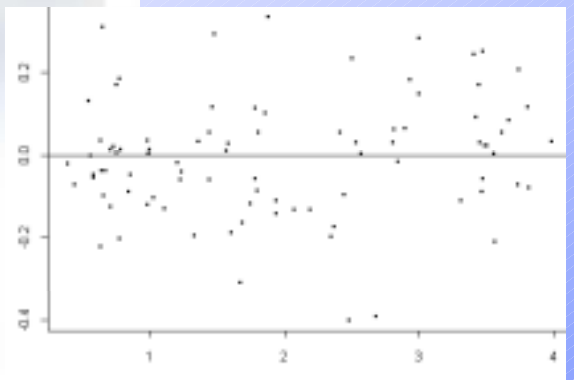
(2)检验方法

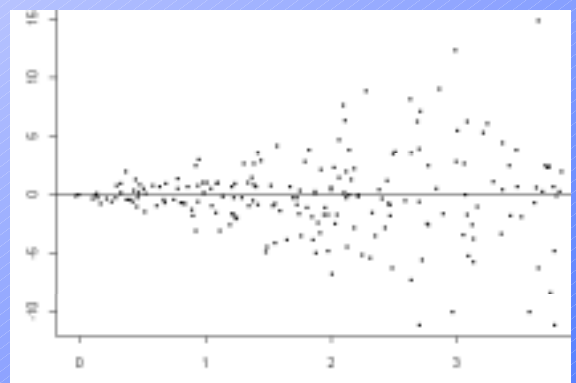
- t检验
- F检验(一元回归中,F检验与t检验一致,即: $F=t^2$,两种检验可以相互替代

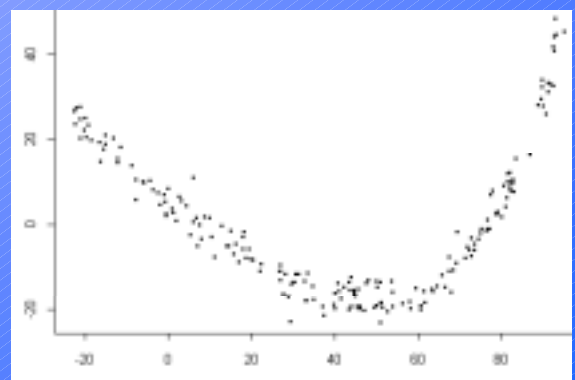
一元线性回归方程的检验

(三)残差分析

- 残差序列的正态检验
- 残差序列的随机性检验
- 残差序列的独立性检验
- 异常值诊断







线性回归方程的预测

(一)点估计

y_0

(二)区间估计

95%的近似置信区间:

y0-2Sy,y0+2Sy

x0为xi的均值时,预测区间最小,精度最高.x0越远离均值,预测区间越大,精度越低.

一元线性回归分析应用举例

工人和管理人员

- 观察R^2值
- 观察方差分析表
- 观察t检验和F检验的关系
- 能够写出回归方程
- 利用相关分析绘制预测值的置信区间

多元线性回归分析

(一)多元线性回归方程

多元回归方程: $y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$

- β_1、β_2、β_k为偏回归系数。
- β_1表示在其他自变量保持不变的情况下，自变量x_1变动一个单位所引起的因变量y的平均变动

(二)多元线性回归分析的主要问题

- 回归方程的检验
- 自变量筛选
- 多重共线性问题

多元线性回归方程的检验

(一)拟和优度检验

判定系数R^2和调整的$R^{2:}$

(二)回归方程的显著性检验：

(1)目的:检验所有自变量与因变量之间的线性关系是否显著，是否可用线性模型来表示.

(2) 利用F检验,构造F统计量

(三)回归系数的显著性检验

(1)目的:检验每个自变量对因变量的线性影响是否显著.

(2)利用t检验,构造t统计量：

多元线性回归分析中的自变量筛选

(一)自变量筛选的目的

- 多元回归分析引入多个自变量. 如果引入的自变量个数较少,则不能很好的说明因变量的变化;
- 并非自变量引入越多越好.原因:
  - 有些自变量可能对因变量的解释没有贡献
  - 自变量间可能存在较强的线性关系,即:多重共线性. 因而不能全部引入回归方程.

多元线性回归分析中的自变量筛选

(二)自变量筛选法

- 向前筛选法(forward)，是自变量不断进入回归方程的过程.
- 向后筛选法(backward)，是自变量不断剔除出回归方程的过程
- 逐步筛选法(stepwise)，是“向前法”和“向后法”的结合

线性回归分析中的共线性检测

(一)共线性带来的主要问题

(二)共线性诊断

- 自变量的容忍度(tolerance)和方差膨胀因子
- 用特征根刻画自变量的方差

多元线性回归分析应用举例

(一)根据10个市场区在特定周内某产品的销售额、广告费、人口密度数据,建立销售额的预测模型

- 所有自变量强行进入方程(方程存在作用不显著的自变量)
  - 观察方差分析表
  - 观察t检验
  - 观察回归方程标准误差和R2
- 逐步回归,与上述参数进行比较(虽然误差增大)

多元线性回归分析应用举例

(二)根据各年国民收入等指标的数据，建立回归模

- 各种自变量筛选方法
  - 观察自变量进入情况
- 观察回归方程的各种检验参数
  - 调整R^2、回归方程标准误差 、F检验、t检验

多元线性回归分析应用举例

(三)分析妇女的年龄、文化程度和居住地区对所生子女数的影响

- 特点:自变量中含定序变量.
- 方法:采用取值为0或1的虚拟变量
- 参照类:不设虚拟变量明确表示的类别(例:文化程度中取文盲半文盲为参照类;居住地中农村为参照类)
- 虚拟变量回归系数的解释，表示该类别与参照类上均值的差异

曲线估计(curve estimate)

(一)目的:

在一元回归分析或时间序列中,因变量与自变量(时间)之间的关系不呈线性关系,但通过适当处理,可以转化为线性模型.可进行曲线估计.

(二)曲线估计的常用模型:

- $y=b_0+b_1t$ (线性拟和linear)
- $y=b_0+b_1t+b_2t^2$ (二次曲线quadratic)
- $y=b_0+b_1t+b_2t^2+b_3t^3$ (三次曲线cubic)

t为时间,也可为某一自变量.

曲线估计应用举例

- 利用总产值和邮件量的样本数据,建立总产值关于邮件量的回归方程
  - 二次曲线和三次曲线
  - 趋势外推预测

Logistic模型

- 问题提出
  - 因变量为分类变量，分类变量间的差距是不等距的
  - 如果因变量表示事件发生的概率率，通常与自变量之间不存在线性关系
  - 不能保证在自变量的各种组合下，因变量的取值仍限制在0~1内

Logistic模型

- 模型简介

$$P = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m$$

$$\log it(p) = \ln \frac{p}{1-p}$$

$$\log it(p) = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m$$

其中：P为因变量取值为1的概率，P/(1-P)称为发生比

Logistic模型

- 模型回归系数的意义
  - 回归系数表示当其他自变量取值保持不变时，该自变量取值增加一个单位引起发生比（OR）自然对数值的变化量。
  - 用发生比($\exp(b_i)$)测量自变量x_i变化对发生概率的影响程度

Logistic模型

- 模型的参数检验
  - 对于整体模型的检验
    - 似然函数值(Likelihood)越大越好
    - -2log likelihood(-2LL)越小意味着回归方程的似然值越大，模型的拟合程度越高；反之；
    - 以截距模型或上一个模型为参照，评价自变量对因变量的解释贡献，即：

    $-2LL_0-(-2LL_x)$，服从卡方分布，P
  - 回归系数的检验
    - Wald统计量
    - 当回归系数很大时，Wald统计量存在一定偏差

Logistic模型

- 分析实例
  - 家庭年收入、居住地区类型与是否购买电脑
    - Enter方式
    - 自变量的交互作用。共线性
    - 模型预测效果
    - Forward:Conditional原则
    - 虚拟变量的处理