线性回归模型

@文彤老师



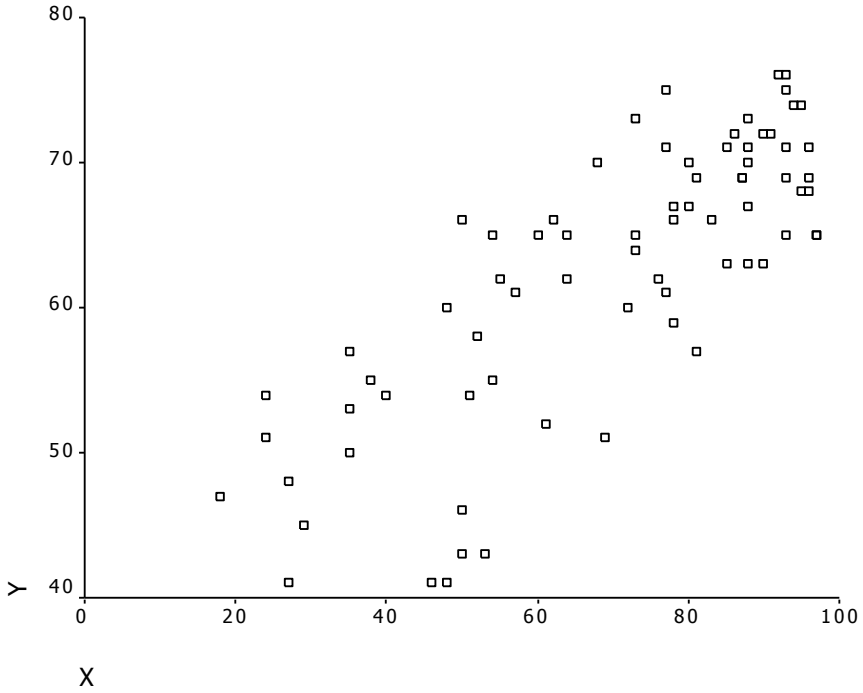
相关分析

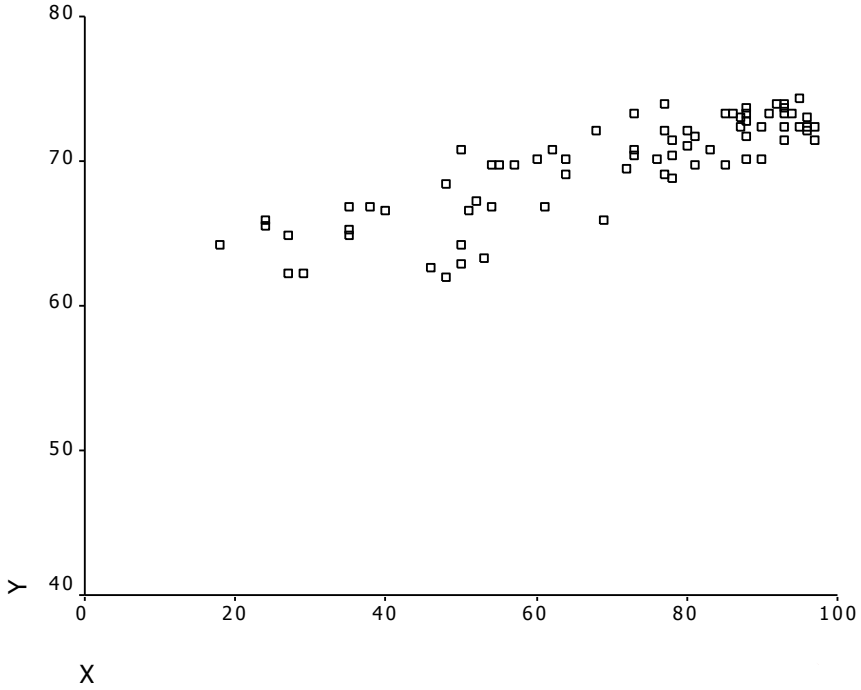
- 任意多个变量都可以考虑相关问题
- 任意测量尺度的变量都可以测量相关强度
- 常用术语
  - 直线相关
    - 两变量呈线性共同增大
    - 呈线性一增一减
  - 曲线相关
    - 两变量存在相关趋势
    - 并非线性，而是呈各种可能的曲线趋势
  - 正相关与负相关
  - 完全相关

相关分析

- 分析过程介绍
  - Bivariate过程
    - 进行两个/多个变量间的参数/非参数相关分析
    - 如果是多个变量，则给出两两相关的分析结果
  - Partial过程
    - 对其他变量进行控制
    - 输出控制其他变量影响后的相关系数
    - 这种分析思想和协方差分析非常类似
  - 典型相关分析

相关分析和回归分析的关系





回归分析概述

- 研究一个连续性变量（因变量）的取值随着其它变量（自变量）的数值变化而变化的趋势
- 通过回归方程解释两变量之间的关系显的更为精确，可以计算出自变量改变一个单位时因变量平均改变的单位数量，这是相关分析无法做到的
- 除了描述两变量的关系以外，通过回归方程还可以进行预测和控制，这在实际工作中尤为重要

回归分析概述

- 回归分析假定自变量对因变量的影响强度是始终保持不变的，如公式所示：

$$\hat{y} = a + bx$$

- 对于因变量的预测值可以被分解成两部分：
  - 常量（constant）：x取值为零时y的平均估计量，可以被看成是一个基线水平
  - 回归部分：它刻画因变量Y的取值中，由因变量Y与自变量X的线性关系所决定的部分，即可以由X直接估计的部分

回归分析概述

$$\hat{y} = a + bx$$

- Ŷ：y的估计值（所估计的平均水平），表示给定自变量的取值时，根据公式算得的y的估计值
- a：常数项，表示自变量取值均为0时因变量的平均水平，即回归直线在y轴上的截距
  - 多数情况下没有实际意义，研究者也不关心
- b：回归系数，在多变量回归中也称偏回归系数。自变量x 改变一个单位，y估计值的改变量。即回归直线的斜率

回归分析概述

- 估计值和每一个实测值之间的差被称为残差。它刻画了因变量y除了自变量x以外的其它所有未进入该模型，或未知但可能与y有关的随机和非随机因素共同引起的变异，即不能由x直接估计的部分。
- 为了方程可以得到估计，我们往往假定εi服从正态分布$N(0,\sigma^2)$。

$$y_i = a + bx + e_i \qquad e_i \sim N(0, \sigma^2)$$

案例：销量影响因素分析

- 某专门面向年轻人制作肖像的公司计划在国内再开设几家分店，收集了目前已开设的分店的销售数据(Y，万元)及分店所在城市的16岁以下人数(X1，万人)、人均可支配收入(X2，元)，数据见reg.sav。试进行统计分析。
  - 实际上拟合的模型如下：

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

$$y_i = \hat{y} + e_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i$$

模型适用条件

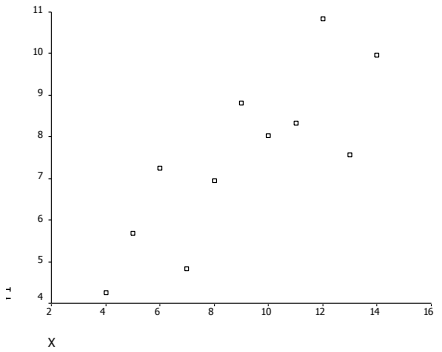
- 线性趋势
- 独立性
- 正态性
- 方差齐性
  - 如果只是探讨自变量与因变量间的关系，则后两个条件可以适当放宽
- 样本量
  - 根据经验，记录数应当在希望分析的自变量数的20倍以上为宜。
  - 实质上样本量和模型的决定系数有关，可通过迭代的方法进行计算

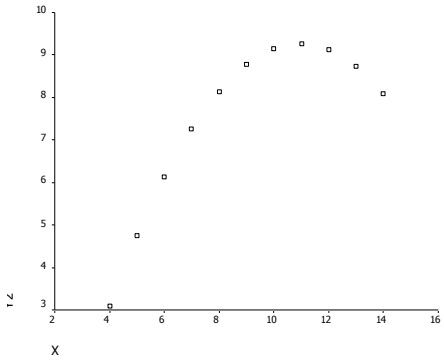
常用指标

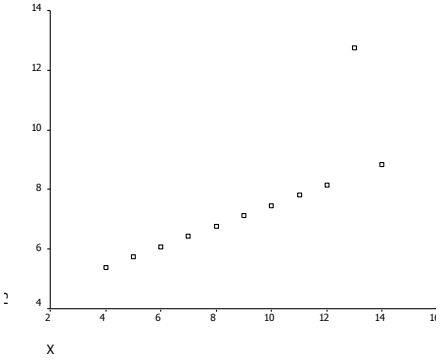
- 偏回归系数
  - 相应的自变量上升一个单位时，因变量取值的变动情况，即自变量对因变量的影响程度。
- 标化偏回归系数：量纲问题
- 决定系数
  - 相应的相关系数的平方，用R^2表示，它反映因变量y的全部变异中能够通过回归关系被自变量解释的比例。

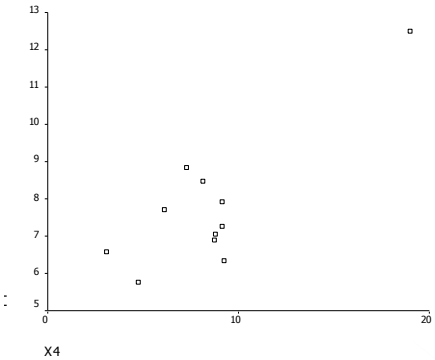
线性回归模型简介

- 分析步骤
  - 做出散点图，观察变量间的趋势







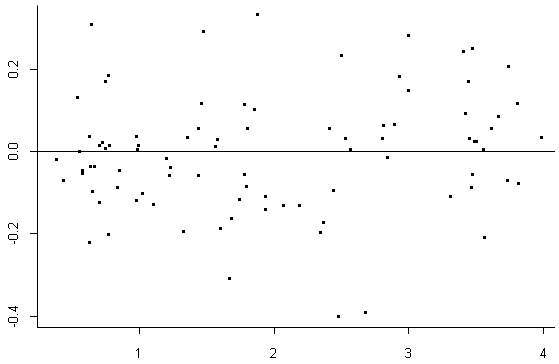


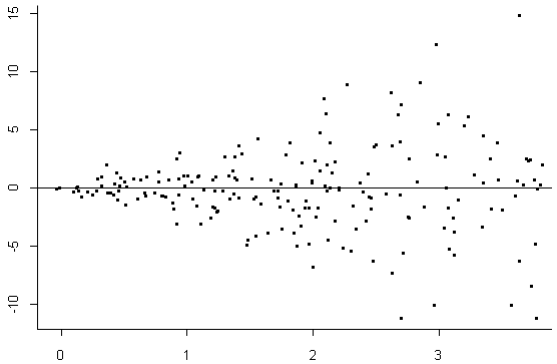
线性回归模型简介

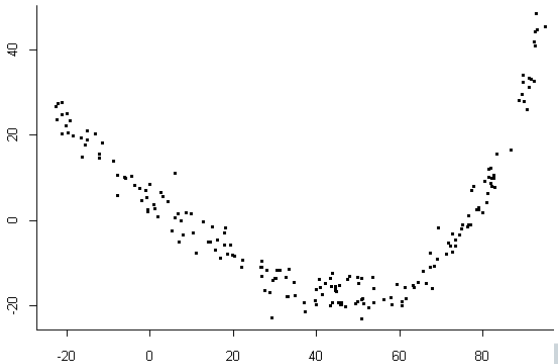
- 分析步骤
  - 考察数据的分布，进行必要的预处理。即分析变量的正态性、方差齐等问题
  - 进行直线回归分析
  - 残差分析
    - 残差间是否独立（Durbin-Watson检验）
    - 残差分布是否为正态（图形或统计量）

线性回归模型简介

- 分析步骤
  - 残差分析







  - 强影响点的诊断
  - 多重共线性问题的判断
    - 这两个步骤和残差分析往往混在一起，难以完全分出先后

案例：固体垃圾排放量与土地种类的关系

- 本例来自Golueke and McGauhey1970年对美国40个城市的固体垃圾排放量(吨)的调查资料，所关心的问题是不同种类土地使用面积（单位，英亩）与固体垃圾排放量之间的关系。可能的影响因素有：indust（工业区土地面积的大小）、metals（金属制造企业用地面积）、trucks（运输及批发商业用地面积）、retail（零售业用地面积）、restrnts（餐馆与宾馆用地面积）。试作逐步回归分析。数据库为WASTE.sav。

多变量的筛选策略

- 较稳妥的方式
  - 单自变量回归模型，筛掉那些显然无关联的候选变量
  - 尝试建立多自变量模型，可手动、也可利用自动筛选方法，但使用后者时要谨慎
  - 多自变量和单自变量模型结果相矛盾时，以前者为准
  - 结果不符合专业知识时，尽量寻找原因

回归分析衍生方法

@文彤老师

曲线拟合过程

- 方法简介
  - 直线关系毕竟是较少数的情形，当因变量和自变量呈曲线关系时：
    - 有明确的公式：利用变量变换将曲线直线化，然后加以拟合
    - 关系不明：基于图形观察，拟合可能的曲线，从中挑选出最为合适的一个
  - 具体拟合方法：根据所选择的公式，将自变量和因变量进行变量变换，然后按照直线回归的方式进行拟合
  - 可拟合的曲线种类
    - 高次方曲线：一、二、三次方曲线
    - 指数、对数、幂曲线
    - 特殊类型曲线：S形曲线、生长曲线等

曲线拟合过程

- 案例：通风时间和毒物浓度的曲线方程
  - 根据文献资料，随着通风时间的增加，密闭空间内污染物的浓度应当呈指数方程下降。现考察某通风设备的换气效果，在室内放置了某种挥发性物质（模拟毒物），待其充分分散到室内空气中后开始通风，每一分钟测量一次室内空气中的毒物浓度，请建立时间与空气中毒物浓度的指数方程。curve.sav
    - 已有明确的方程$y=a\times e^{bx}$，按此拟合即可。
    - 等价于先进行变量变换，然后拟合直线方程。

加权最小二乘法

- 所针对的问题：方差齐性被违反
  - 因变量的变异随着某些指标的改变而改变
    - 以地区为观察单位调查某种事物的发生率
    - 研究通货膨胀和失业率对股票价格的影响
      - 高价股票的波动一般都会大于低价股票
    - 需要人为调控各案例在回归中的重要性
  - 解决办法
    - 根据用户提供的可能预测因变量变异大小的指标，在拟合时对变异较小（即测量更精确）的测量值赋予较大的权重

加权最小二乘法

- 案例：不等量样品数据的回归方程
  - 实验中收集得15对数据，每对数据都是将n份样品混合后测得的平均结果，但各对数据的n大小不等，试求出X对Y的直线回归方程。wls.sav
    - 加权后的决定系数基本上都是低于原模型的

岭回归分析简介

- 一种专门用于共线性数据分析的有偏估计方法
  - 有偏意味着对数据信息有所取舍
  - 通过丢弃部分信息，以得到更为稳定的分析结果
  - 实际上是一种改良的最小二乘法
  - 由于是有偏估计，统计检验已经居次要地位，故一般不再给出

岭回归分析简介

- 程序方式调用

```
INCLUDE 'SPSS所在路径\Ridge Regression.sps'.
ridgereg   enter = 自变量列表
           /dep = 应变量名
           /start= K值起始值，默认为0
           /stop= K值终止值，默认为1
           /inc= K值搜索步长，默认为0.05
           /k= 允许搜索的K值个数，默认为999 .
```

岭回归分析简介

- 案例：用外形指标推测胎儿周龄
  - 现测得22例胎儿的身长、头围、体重和胎儿受精周龄，具体数据见文件ridgereg.sav。研究者希望能建立由前三个外形指标推测胎儿周龄的回归方程

```
INCLUDE 'd:\spsswin\Ridge Regression.sps'.
ridgereg enter = long touwei weight        列出方程中的自变量
/dep = y                                   列出方程的应变量
/ inc = 0.01 .                             要求以 0.01 为步长搜索 k 值
```

最优尺度回归

- 所针对的问题：测量尺度非等距
  - 变量为无序多分类时，类别间的差异如何较难探索
  - 变量为有序多分类时，类别间的距离不一定相等，直接作为连续变量分析不妥
  - 即使变量均为连续型变量，但其联系有可能为某种曲线，直接按照线性结构来拟合也不合适

最优尺度回归

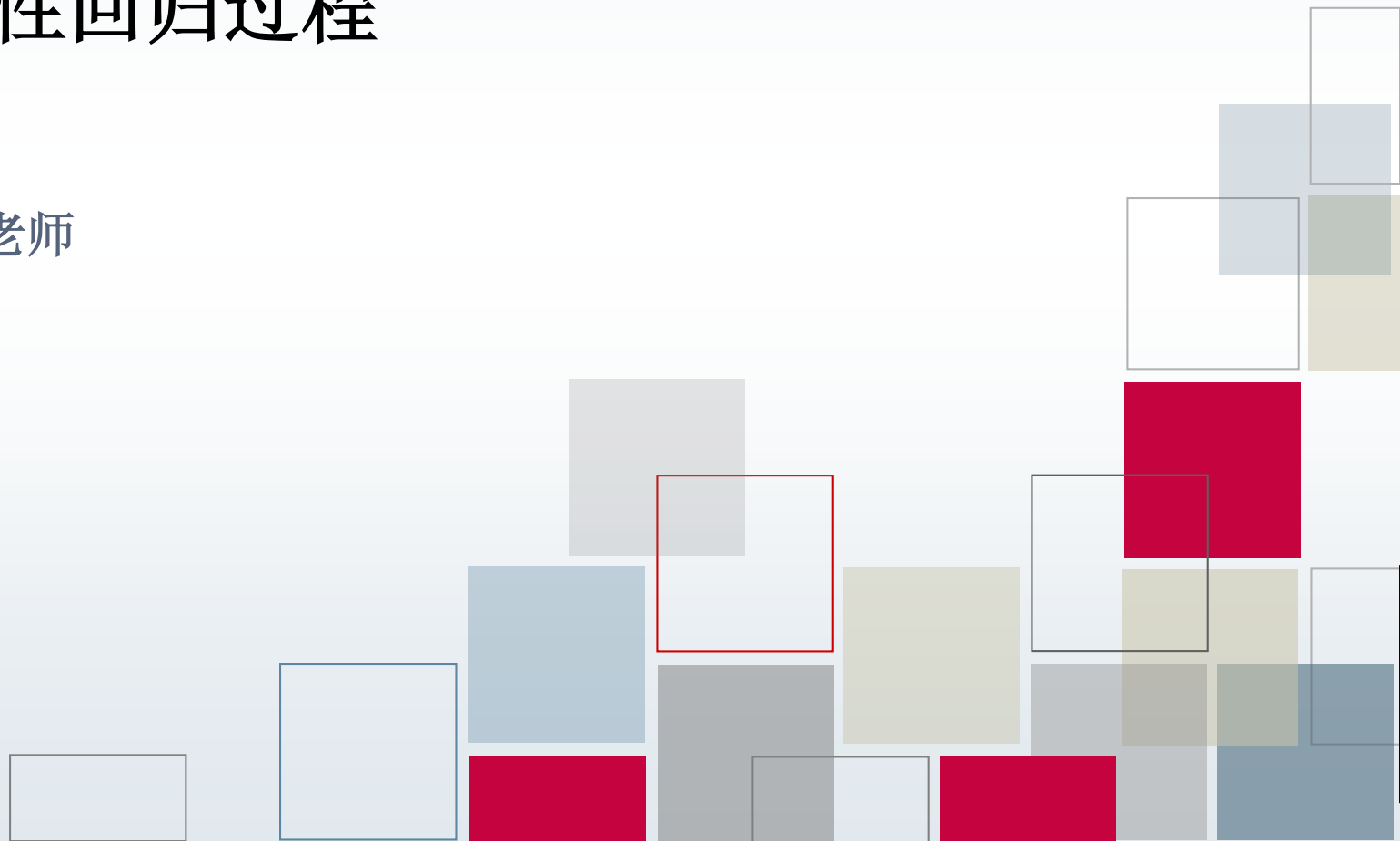
- 解决办法
  - 根据数据情况进行迭代搜索，找到适当的变换方法对原始分类变量进行转换，将原始变量一律转换为连续性评分，然后再进行方程拟合
    - 分类变量越多优势越明显
  - 从实用的角度出发，该方法可以被作为一种探索性方法使用

最优尺度回归

- 案例：生育子女数的回归模型
  - 现收集了一批妇女的曾生子女数、年龄、居住地类别（1：城市，2：农村）、受教育程度（1~5分别代表文盲半文盲、小学、初中、高中、大学及以上），请建立后三个变量对曾生子女数的回归模型，数据见child.sav。
    - 用此方法来探索一下受教育程度对因变量的影响趋势

非线性回归过程

@文彤老师



曲线拟合过程的局限

- 只能分析一个自变量
- 变量变换的局限
  - 有的公式根本无法进行变换，如复杂的等式，或者无简单解的积分方程
  - 当变换后，变量的数值分布状况已经改变，此时根据最小二乘法得到的最优解可能在原变量分布状况下并非最优

非线性回归过程的优势

- 它采用迭代方法对用户设置的各种复杂曲线模型进行拟合
  - 迭代方法往往意味着结果较为稳定
- 将残差的定义从最小二乘法向外大大扩展
  - 这意味着误差测量手段的大大丰富
    - 最小一乘法、加权最小二乘法、自回归模型等
- 为用户提供了极为强大的分析能力
  - 特别适用于实验室数据的分析

非线性回归过程简介

- 案例：毒物通风数据
  - 在曲线拟合过程中，给出的解实际上是变量变换后线性回归方程的最优解
  - 使用非线性回归拟合时，给出的解为原始变量状况下的最优解
    - 即散点离曲线距离的平方之和为最小
    - 此时的决定系数一般均高于曲线拟合过程

案例：自定义损失函数

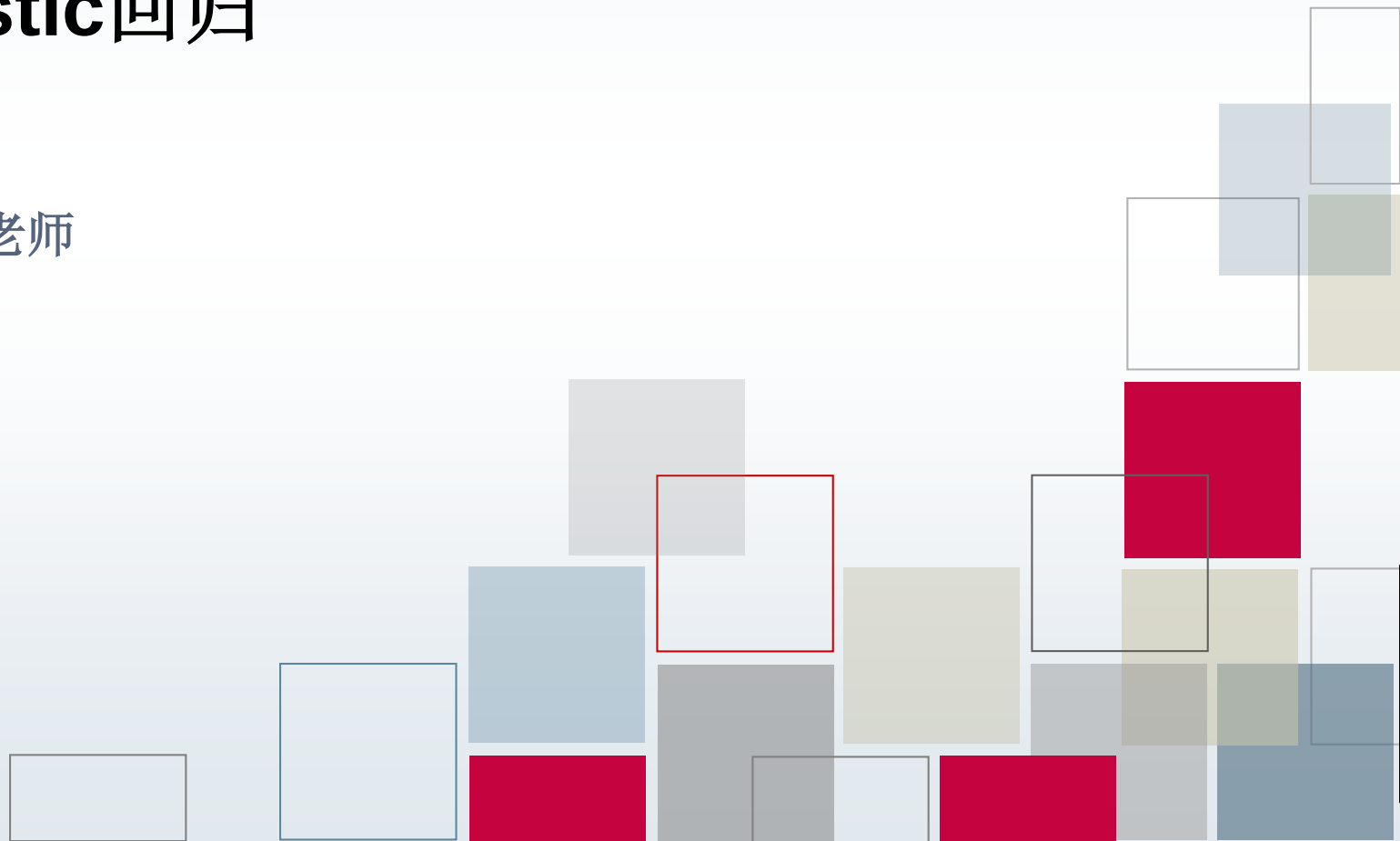
- 某公司生产的产品其成本主要受两种原材料的影响，为及时调整生产，协调库存，现收集了一批产品产量与相应生产中两种原材料消耗量的数据，见文件nlin2.sav。请就此建立原材料消耗量与产量（因变量）间的回归方程

非线性回归过程简介

- 参数初始值的设定技巧
  - 如果可变为线性，可以先拟合线性方程，将此结果作为初始值
  - 如果方程可解，则代入若干样本值，解出近似取值作为初值
  - 先拟合较简单的雏形，将结果作为初始值
  - 否则，多尝试几种初始值，观察结果

logistic回归

@文彤老师



模型简介

- 基于线性回归模型发展而来
  - 线性回归研究的是连续性因变量与自变量之间的关系
- 有的时候因变量为分类变量，需要研究该分类变量与一组自变量之间的关系
  - 以治疗效果为因变量，结局为治愈/未治愈
  - 如果使用新的宣传方式，决定戒烟的概率是否更高？

模型简介

$$\hat{Y} = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m$$

$$\hat{P} = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m$$

- 发生率P为因变量，它与自变量之间通常不存在线性关系
- 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

模型简介

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

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

- 由于因变量为二分类，所以误差项服从二项分布，而不是正态分布
  - 因此，常用的最小二乘法也不再适用

模型用途

- 影响因素分析，求出哪些自变量对因变量发生概率有影响。并计算各自变量对因变量的比数比
- 作为判别分析方法，来估计各种自变量组合条件下因变量各类别的发生概率，从而对结局进行预测。
  - 该模型在结果上等价于判别分析

模型简介

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

- α是常数项，表示自变量取值全为0时，比数（Y=1与Y=0的概率之比）的自然对数值
- Beta为logistic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比数比（OR）自然对数值的变化量

案例：低出生体重儿影响因素

- Hosmer和Lemeshow于1989年研究了低出生体重婴儿的影响因素，数据见文件logistic_step.sav。
- 结果变量为是否娩出低出生体重儿(变量名为LOW，1，低出生体重，即婴儿出生体重<2500克、0，非低出生体重)，考虑的影响(自变量)有：
  - 产妇妊娠前体重(lwt，磅)
  - 产妇年龄(age，岁)
  - 产妇在妊娠期间是否吸烟(smoke，0=未吸、1=吸烟)
  - 本次妊娠前早产次数(ptl，次)
  - 是否患有高血压(ht，0=未患、1=患病)
  - 子宫对按摩、催产素等刺激引起收缩的应激性(ui，0=无、1=有)
  - 妊娠前三个月社区医生随访次数(ftv，次)
  - 种族(race，1=白人、2=黑人、3=其他民族)

模型中用到的检验方法

- Walds检验：基于标准误估计值的单变量检验
  - 他没有考虑其他因素的综合作用，当因素间存在共线性时结果不可靠
  - 故在筛选变量时，用Walds法应慎重
- 似然比检验：直接对两个模型进行的比较
  - 当模型较为复杂时，建议使用似然比检验进行变量的筛选工作，以及模型间优劣的比较
  - 两模型-2对数似然值之差即为似然比统计量，自由度亦为两模型参数个数之差
- 比分检验
  - 考虑在已有模型基础上引入新变量之后模型效果是否发生改变

哑变量

- 回归系数*b*表示其它自变量不变，*x*每改变一个单位时，所预测的*y*的平均变化量
- 当x为连续性/二分类变量时这样没有问题
- 当x为多分类变量时就不太合适了
  - 无序多分类：民族，各族之间不存在大小问题
  - 有序多分类：家庭收入分为高、中、低三档，它们之间的差距无法准确衡量
    - 强行规定为等距显然可能引入更大的误差

哑变量

- 在以上这些情况时，我们就必须将原始的多分类变量转化为数个哑变量（Dummy Variable），每个哑变量只代表某两个级别或若干个级别间的差异，这样得到的回归结果才能有明确而合理的实际意义。
  - 注意：哑变量必须同进同出，否则含义可能改变

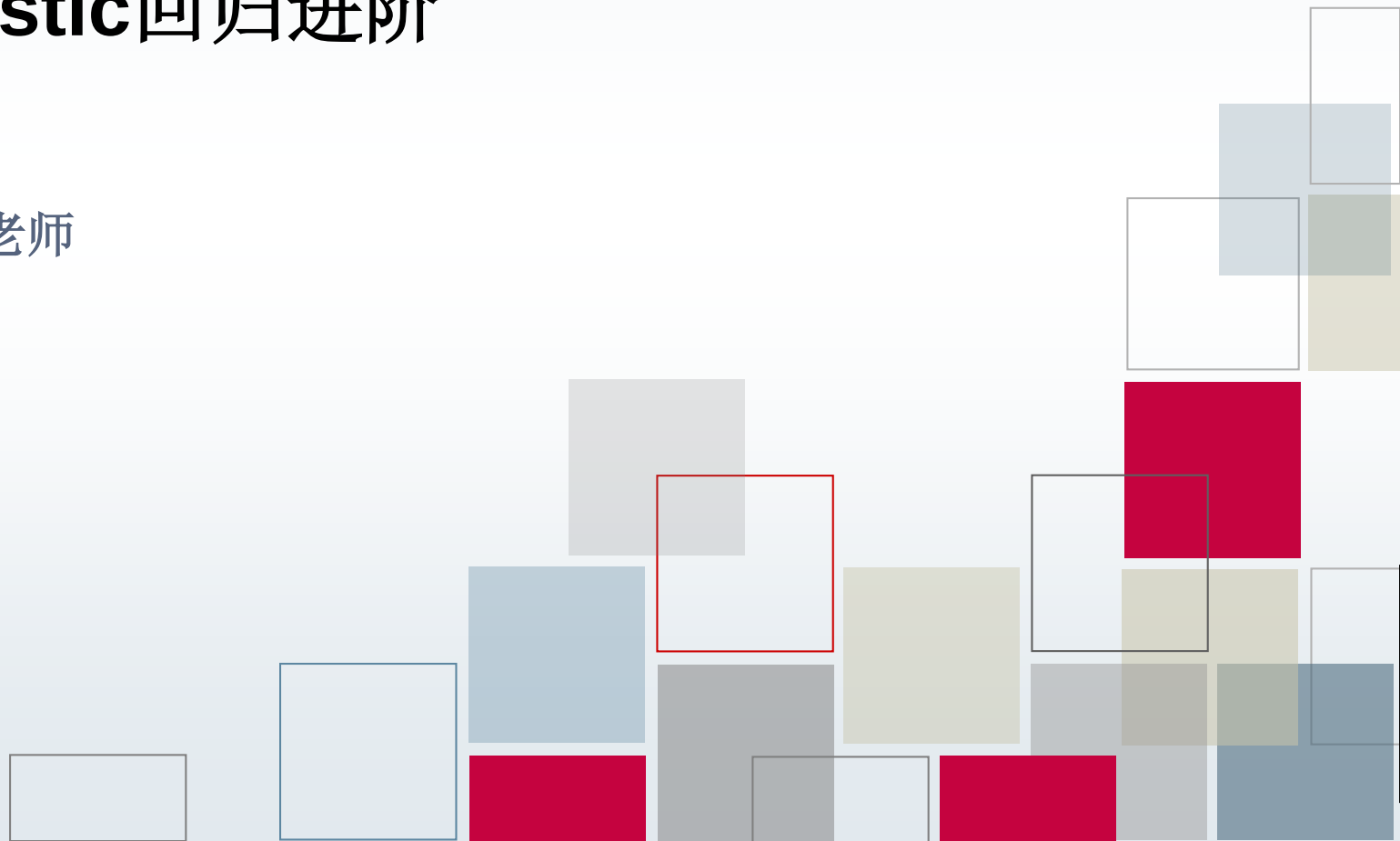
哑变量

- O型是作为对比水平（基础水平），而哑变量V1、V2、V3分别代表了A型、B型、AB型和O型相比的系数

	O型	A型	B型	AB型
V1	0	1	0	0
V2	0	0	1	0
V3	0	0	0	1

Logistic回归进阶

@文彤老师



无序多分类Logistic回归模型

- 研究问题
  - 病例—对照研究中设立一组病例和多组对照，需要分析暴露是否和患病有关，则结局变量为无序三分类，应当使用该模型加以分析。
    - 病例
    - 医院对照
    - 健康人群对照

无序多分类Logistic回归模型

- 因变量为无序多分类
- 除一个对照水平外，以每一分类与对照水平作比较
- 例如结果变量有三个水平：a、b、c，如果以a为参照水平，就可以得到两个Logistic函数，一个是b与a相比，另一个是c与a相比
- 同时应当有：Pa+Pb+Pc=1

无序多分类Logistic回归模型

- 模型简介

$$\log itP_a = \ln\left[\frac{P_a}{P_a}\right] = \ln 1 = 0$$

$$\log itP_b = \ln\left[\frac{P(Y=b \mid X)}{P(Y=a \mid X)}\right] = \alpha_b + \beta_{11}X_1 + \cdots + \beta_{1p}X_p$$

$$\log itP_c = \ln\left[\frac{P(Y=c \mid X)}{P(Y=a \mid X)}\right] = \alpha_c + \beta_{21}X_1 + \cdots + \beta_{2p}X_p$$

无序多分类Logistic回归模型

- 案例：不同背景人群的选举倾向
  - 老布什、克林顿、佩罗在1992年进行的较量，数据来自SPSS自带的vote.sav。
    - pres92，所欲选的总统候选人；
    - age，年龄；
    - agecat，年龄分组；
    - educ，受教育年数；
    - degree，最高学历；
    - sex，性别。

有序多分类Logistic回归模型

- 相关问题
  - 所测量的结局变量为等级，或者数量较少的评分（如1～5分）
    - 疗效：痊愈、显效、好转、无效
  - 单变量分析使用秩和检验即可，如果进行多变量分析，简单的按照连续变量来处理可能不合适

有序多分类Logistic回归模型

- 多分类有序因变量的资料，分类水平大于2且水平之间有等级关系。
- 拟合水平数-1个logit模型，称为累加logit模型（Cumulative logits model）。
- 例如对一个四分类有序变量，即应当同时拟合以下三个模型：π1、π2、π3分别为因变量取第一类、第二类、第三类时的概率，而第四类则作为用于对比的基础水平。

有序多分类Logistic回归模型

- 模型简介

$$\log it_1 = \log(\frac{\pi_1}{1-\pi_1}) = \alpha_1 + \beta' X$$

$$\log it_2 = \log(\frac{\pi_1 + \pi_2}{1-\pi_1-\pi_2}) = \alpha_2 + \beta' X$$

$$\log it_3 = \log(\frac{\pi_1 + \pi_2 + \pi_3}{1-\pi_1-\pi_2-\pi_3}) = \alpha_3 + \beta' X$$

有序多分类Logistic回归模型

- 可见，这种模型实际上是依次将因变量划分为两个等级，不管模型中因变量的分割点在什么位置，模型中各自变量的系数β都保持不变，所改变的只是常数项α。此时求出的OR值是自变量每改变一个单位，因变量提高一个及一个以上等级的比数比。
- 这种假设看似复杂，但大量实践证明，它是符合多数实际情况的。

案例：工作满意度影响因素分析

对某地人群调查其对所从事的工作是否满意，可能的影响因素有：年龄、性别、年收入水平、文化程度。变量的赋值情况见下表，数据见文件 satisfy.sav。试进行统计分析。

Satis(满意度)	①满意②态度中立③不满意
Age(年龄)	岁
Gender(性别)	f 女性、m 男性
Inccat(年收入水平)	①2.5 万以下②2.5 万～③5.0 万～④7.5 万及以上
Edu(文化程度)	①初中及以下②高中③大专④大学⑤研究生

- 各分割点的平行性检验
- 如果平行性检验未能过关
  - 换用其他连接函数
  - 退回去使用无序logistic模型