多元线性回归
(Multiple Linear Regression Analysis)

王丽
流行病学与卫生统计学系

多变量分析方法的提出

- 流行病学的一个重要应用是探索病因或危险因素（包括识别和处理混杂因素及效应修饰因素）。

- 单变量（因素）分析：
  - 分析单一特异性因素引起的健康危害或疾病或其它结局效应
  - 难以处理多因素引起的疾病

- 多变量（因素）引起的疾病的病因研究
  - 研究设计阶段：匹配
  - 统计分析阶段
    - 分层分析
    - 多变量分析

- 分层分析是将可能对结局产生影响的变量（外源性变量或混杂变量），按其不同属性分层，再在每层内分析主要变量与结局的联系
- 研究的变量数目（2或3个）较少时，分层分析方法完全适用。

分层分析的例子

- 口服避孕药与心肌梗死病例对照研究

	服OC	未服OC	计	
MI	39	114	153	$_cOR = 2.19$
对照	24	154	178	
计	63	268	331	

- 如果怀疑年龄有混杂作用，按年龄分层再进行分析。

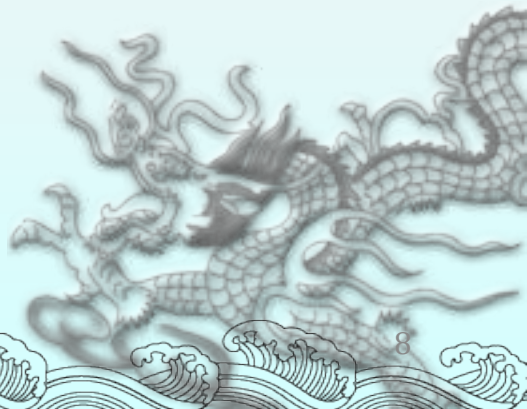
口服避孕药与心肌梗死病例对照研究

按年龄分层后的结果

	< 40岁			≥ 40岁		
	OC（+）	OC（−）	计	OC（+）	OC（−）	计
MI	21（a_1）	26（b_1）	47（m_{11}）	18（a_2）	88（b_2）	106（m_{12}）
对照	17（c_1）	59（d_1）	76（m_{01}）	7（c_2）	95（d_2）	102（m_{02}）
计	38（n_{11}）	85（n_{01}）	123（n_1）	25（n_{12}）	183（n_{02}）	208（n_2）
	$OR_1 = 2.80$			$OR_2 = 2.78$		
	${}_cOR/OR_1 = 0.78$			${}_cOR/OR_2 = 0.78$		

常用的多变量分析方法

- 协方差分析
- 多元线性回归
- logistic回归
- 比例风险回归（Cox回归）
- 多重（偏）相关分析
- 主成分分析
- 因子分析
- 聚类分析



统计学处理方法的选择

- 两个或以上自变量资料的统计学分析方法

自变量	因变量	统计方法
属性（有混杂）	数值	协方差分析
属性或数值	数值	多元回归
属性或数值	属性（二分）	logistic回归
数值或属性	二分（属性）	发生的风险（有截缩）Cox（比例风险）回归
属性	属性	对数-线性
属性或数值	属性（多分）	判别分析
数值	—	因子分析或聚类分析

出生体重危险因素研究

- ID　　编号
- LOW　　出生低体重(bwt<2500=1,>=2500=0)
- AGE　　母亲年龄(岁)
- LWT　　母亲末次月经时的体重
- RACE　　种族：1 白种人　2 黄种人　3 黑人
- SMOKE　　吸烟史：　1 吸烟　0 否
- PTL　　早产史：　1　有　　0 无
- HT　　妊娠高血压：　1 有　　0 无
- UI　　频繁宫缩：1 有　　0 无
- FTV　　产前访视次数
- BWT　　出生体重(克)

探讨的问题

- 婴儿的出生低体重（low)是否与母亲的种族(race)有关？
- 黑人母亲和非黑人母亲的婴儿出生体重是否有显著性差别？
- 黑人、白人及黄种人母亲，其婴儿的出生体重是否有显著性差别？
- 母亲的年龄、吸烟史、既往早产史、妊娠高血压史，频繁宫缩史是否分别与婴儿的出生体重有关？
- 母亲的年龄、吸烟史、既往早产史、妊娠高血压史，频繁宫缩史哪些因素与婴儿的出生体重有关？
- 在控制了年龄、目前吸烟史、既往早产史、妊娠高血压史，频繁宫缩史之后，婴儿的出生体重是否依旧与种族有关？
- 亲的年龄、吸烟史、既往早产史、妊娠高血压史，频繁宫缩史哪些因素与婴儿的出生低体重有关？

多变量线性回归分析

一、概念

- 回归是研究变量与变量之间关系的一种手段，通过回归方程表达变量与变量之间的一种依存关系。

- 当研究变量之间的线性关系时就是直线回归（linear regression）

- 如：UCSF大学的妇产科学及生殖研究所收集1980年-1990年在该生殖中心出生的婴儿及其母亲的资料。
- 母亲的信息：怀孕时的年龄、吸烟史、怀孕前的体重、早产史、是否有妊娠高血压、怀孕期间是否发生频繁宫缩、产前接受的访试次数等。
- 新生儿的信息：出生时的体重
- 要回答的问题：用回归方程定量的刻画一个新生儿出生体重（因变量Y)与母亲孕期的多个自变量X1，X2，.....间的线性依存关系

出生体重危险因素研究数据库字段注释

- 变量名　字段注释
- ID　编号
- BWT　出生体重（克）
- AGE　母亲年龄（岁）
- LWT　母亲末次月经时的体重（磅）
- RACE　种族：1白种人 2黄种人 3黑人
- SMOKE　吸烟史：1吸烟 0否
- PTL　早产史：1有 0否
- HT　妊娠高血压：1有 0否
- UI　频繁宫缩：1有 0否
- FTV　产前访试次数：1一次 2二次及以上 0无

- 设有p个自变量X_1，X_2，...,X_p,一个因变量Y，以及一份由n个个体构成的随机样本（x_{i1}，x_{i2}，...，x_{ip}，y_i），i=1，2，...,n

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

a：回归方程常数项

b_p：偏回归系数，指其它自变量固定的条件下，某自变量Xp每改变一个单位时，因变量Y的平均变化量。

$\hat{Y}$ ：在给定自变量取值条件下y的均数的一个点估计。

新生儿出生体重与母亲怀孕时相关因素的关系：

对每一个新生儿而言：

$Y_i=b0+b1*age_i+b2*smoke_i+b3*lwt_i+...$

根据所有新生儿及其母亲的观测值，可以得到新生儿出生体重与母亲相关因素的回归方程：

$\hat{Y_i}=b0+b1*age_i+b2*smoke_i+b3*lwt_i+....$

残差：

二、线性回归方程需满足的条件

（一） LINE 原则

- **L（linear）：**

  自变量和因变量呈线性关系；

- **I （independence）：**

  某xi值改变对y的影响与另一xi的水平无关；

  y呈独立性，即任一个体的y值对另一个体的y值不提供任何信息；

- **N（normality）：**

  xi分别取某定值时得到的一组y值呈正态分布；

- **E （equal variance）：**

  各y值的方差相等，即各xi取不同值时y的不同分布服从方差齐性，即其方差为常数

（二）因变量的选择

因变量必须是定量指标，同时必须满足以上关于线性回归的条件要求，即**LINE**。

（三）自变量的选择

- 对于自变量没有强制性要求，但自变量和因变量之间必须是线性关系。

- 自变量可以为定量指标、定性指标以及等级变量中的任何一种。

- 如果自变量为定量指标：

（1）同时自变量与因变量之间为线性关系，则可以直接以原变量的形式进入分析；

（2）如果自变量与因变量之间为非线性关系，则需做适当转换，如x2，log（x），ex等。

- 自变量为定性或等级指标：

不需要做自变量与因变量的线性关系检验

- 自变量为定性指标：
  - 为二分类变量，常用0，1或1，2表示。如x为性别指标，0代表女性，1代表男性，回归方程中对应的回归系数b表示男性比女性的y平均多b。
  - 为多分类指标，需要专业判定指标的意义
    - 等级变量：直接带入分析
    - 定性变量：亚变量（dummy）

亚（哑）变量的设置

- 引入亚（哑）变量的目的在于区分某个变量的不同属性。
- 当自变量为属性变量，特别是不同属性之间无等级高低之分，为说明不同属性对因变量的影响大小，常需引入亚（哑）变量。

亚变量的设置：例1

- 一项探讨影响新生儿出生体重的研究：
- 因变量即结局变量为新生儿出生时的体重；
- 研究的因素包括母亲怀孕时的年龄、母亲末次月经时的体重、母亲的种族、是否吸烟、是否有过早产史、是否有妊娠高血压、怀孕期间是否发生频繁宫缩、产前接受的访试次数等。

亚变量的设置例1：出生体重危险因素研究

- 变量名　字段注释
- ID　编号
- BWT　出生体重（克）
- AGE　母亲年龄（岁）
- LWT　母亲末次月经时的体重（磅）
- RACE　种族：1白种人 2黄种人 3黑人
- SMOKE　吸烟史：1吸烟 0否
- PTL　早产史：1有　0否
- HT　妊娠高血压：1有 0否
- UI　频繁宫缩：1有 0否
- FTV　产前访试次数：1一次 2二次及以上 0无

亚变量的设置：例1

- 将种族分成白人、黑人和黄种人3种属性，可引入2（= 3–1）个亚变量，分别表示各种族，选择其中之一（例如，白人）作为参照

变量	x_1	x_2	
白人	0	0	（参照）
黑人	1	0	
黄种人	0	1	

亚变量的设置：例2

- Framingham心脏病研究，随访1,406人18年
- 探讨冠心病发生率与年龄、性别、收缩血压关系的多变量线性回归
- 如何处理年龄与冠心病发生率的关系？
  - 连续变量？
  - 其他？

亚变量的设置：例2（续）

- $x_1 = 0$，$x_2 = 0$，$x_3 = 0$，为40～49岁（参照）
- $x_1 = 1$，$x_2 = 0$，$x_3 = 0$，为50～54岁
- $x_2 = 1$，$x_1 = 0$，$x_3 = 0$，为55～59岁
- $x_3 = 1$，$x_1 = 0$，$x_2 = 0$，为60～62岁

年龄（岁）	x_1	x_2	x_3
40～49（参照）	0	0	0
50～54	1	0	0
55～59	0	1	0
60～62	0	0	1

亚变量的设置：例2（续）

- 得到冠心病发生率与年龄、性别、收缩血压关系的多变量线性回归方程如下：
- $y = 0.0613 + 0.0277 \times x_1 + 0.0826 \times x_2 + 0.0845 \times x_3 + 0.1273 \times x_4 + 0.1680 \times x_5$
- 上式中y = log（冠心病发生率/（1-冠心病发生率））

亚变量的设置：例2（续）

- 有时自变量（如年龄）虽然是连续变量，但按其每改变一个单位（一岁），来估计其对因变量的影响很微弱，如将其划分成大小不同的几种属性，并设立亚变量，则可看出不同属性对因变量的影响大小。

- 这种指标分解方法的优点是有助于分清究竟哪种属性对所研究疾病危险性的作用较大，也便于研究因素间的交互作用。

三、线性回归基本SAS程序

不可省略

不可省略，可用交谈式执行

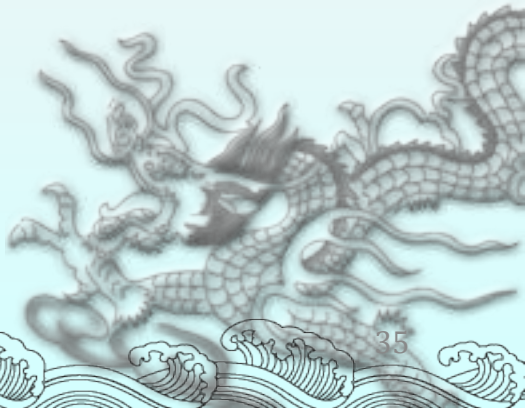
必须在RUN；指令之前

这些指令可放在MODEL指令后的任何一处而且可在交谈式环境下执行

```
PROC REG          选项串;
     MODEL        因变量名称串=自变量名称串 /选项串;
     VAR          变量名称串;
     FREQ         变量名称;
     WEIGHT       变量名称;
     ID           变量名称;
     BY           变量名称串;
     ADD          变量名称串;
     DELETE       变量名称串;
     RESTRICT     等式1，等式2，...;
     TEST         等式1，等式2，.../选项;
     MTEST        等式1，等式2，.../选项串;
     OUTPUT       OUT=输出文件名关键字=变量名称串;
     REWEIGHT     加权条件式 ALLOBS /选项串;
     (或 REWEIGHT STATUSUNDO;)
     REFIT;
     PAINT        强化条件式 ALLOBS /选项串;
     (或 PAINTSTATUSUNDO;)
     PLOT         图形指令串 /选项串;
     PRINT        选项串 ANOVA MODELDATA;
```

PROC REG data=文件名;

调用**REG**过程并指明对哪个文件执行分析，若省略“data= “ ，则SAS会自动找出在本程序之前最后形成的SAS语句。



- **Model 因变量=自变量/选择项；**
- **每次调用REG过程至少要有一个MODEL语句。**
- **MODEL Y=X；　一个应变量对一个自变量的回归**
- **MODEL Y=X1 X2 X3；　一个应变量对多个自变量的回归**
- **MODEL Y1 Y2=X1 X2 X3；　多个应变量对多个自变量的回归**

MODEL语句中的选择项之一：

界定有关参数估计值的有关选项：

- /STB：要求计算模型中各自变量的标准回归系数；
- /CLM：计算出预测值平均数的95%可信区间的上、下限；
- /CLI：计算出各预测值的95%可信区间的上、下限；

多元回归基本SAS程序

```
proc reg;
  model y=x1 x2/stb;
run;
```

四、几个重要的概念

- 偏回归系数：
- 标准回归系数：
- 决定系数：
- 校正复相关系数：
- 剩余标准差：

偏回归系数（b_j）

- 当方程中其他自变量固定时，X_j每改变一个单位，引起Y的平均变化量，也就是说b_j的大小反映了X_j对Y的影响程度。

标准回归系数

- 由于各自变量取值的单位及其离散程度通常不同，所以各量纲不同的回归系数之间不能直接比较大小。为此，需要对偏回归系数进行标准化以消除量纲的影响。
- $$b_j^{'} = b_j \sqrt{\frac{l_{jj}}{l_{YY}}} = b_j \frac{S_j}{S_Y}$$
- 标准回归系数绝对值的大小可用来衡量自变量对应变量Y的贡献大小，以说明各变量在多元回归方程中的重要性。

决定系数（R^2）

- $R^2=SS_{回}/SS_{总}$
- 取值范围在0与1之间，无单位。反映了回归贡献的相对程度，也就是在Y的总变异中回归所能解释的百分比。
- 主要通过决定系数数值的大小来反映回归或相关的实际效果。
- 例如：决定系数=0.9587，说明所求的回归方程能够解释的应变量变异占应变量总变异的95.87%
- 存在的问题：随方程中自变量的增加而加大，即使引入无显著性变量，其值也会略有增加

校正决定系数（R^2_{adj}）

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-m-1}$$

- 其中n为拟合模型的观察单位数；
- m为方程中所含变量个数

剩余标准差

- 扣除各自变量Xj组合对应变量Y的线性关系影响后所剩下的变异。
- 回归估计精度的指标
- 越小回归方程估计误差也越小，估计精度越高。

$$S_{Y,12\ldots m}=\sqrt{\frac{Q}{n-m-1}}$$

- 剩余标准差一般随方程中自变量的增加而减少
- 但若引入某些对应变量Y无显著作用的自变量时，由于回归平方和增加很小，剩余平方和减少很小，但剩余自由度却减少，故求得的剩余标准差反而加大。
- 即方程中增加有显著作用的变量时，R^2_{adj}增加，MSE减少；而方程中引入无显著作用变量时，R^2_{adj}可能减小，MSE反而加大。
- 因此，常以R^2_{adj}越大，MSE越小作为多元回归方程估计效果评价的指标。

五、线性回归方程的评价

（一）评价整个方程在α水准下是否有显著性

- 在SAS软件中，对多元线性回归方程的假设检验采用方差分析进行

$$F = \frac{U/m}{Q/(n-m-1)} \sim F_{\alpha(m,n-m-1)}$$

U：回归平方和，反映由于方程中m个自变量与应变量Y间的线性关系，而使应变量Y变异减小的部分；

m为回归自由度，即方程中所含自变量的个数；

Q为剩余平方和，说明除自变量外，其他随机因素对Y变异的影响；

n-m-1为剩余自由度。

- 如果整个方程在指定的α水准下有显著性意义时，并不说明方程中每个自变量x_j都对y_i有显著性影响。还需对各个自变量的偏回归系数逐个进行检验。

- 但如果整个方程经F检验无显著性，就不必对b_j逐个进行检验。