CDA 统计课程

==========

2.一元线性回归

2.1 最小二乘估计

2.2 线性回归与相关

*相关分析

*偏相关分析

*交互效应模型

2.3 线性回归与方差

*方差分析原理

*方差分析步骤

2.4 数据分析流程

2.1 最小二乘估计

最小二乘技术是最佳估计的常见使用方法（《回归分析》）。

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$$\Rightarrow \quad \frac{\partial D}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial D}{\partial b_1} = -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

$$\Rightarrow \quad n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} (x_i{}^2) = \sum_{i=1}^{n} (x_i y_i)$$

$$\Rightarrow \quad b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)(y_i - \hat{y}_i)}{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2} = \frac{cov(x,y)}{var(x)}$$

==========

2.2 线性回归与相关

（1）相关分析

用于衡量两类现象在发展变化的方向与大小方面存在一定的关联(不包括因果和共变关系)。

类型：有线性相关和非线性相关两种。

一般情况下，如果不做特殊说明，指的就是线性相关。

积差相关公式：

$$r=\sum(x-\bar{x})(y-\bar{y})/\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}$$

注：$\bar{x}$ 表示变量 x 的均值；$\bar{y}$ 表示变量 y 的均值；r 取值范围[-1 1]。

H0：两变量间无直线相关关系。

适用条件：

① 数据间相互独立，包括观测间相互独立与变量间相互独立。

② 两列变量均服从正态分布。

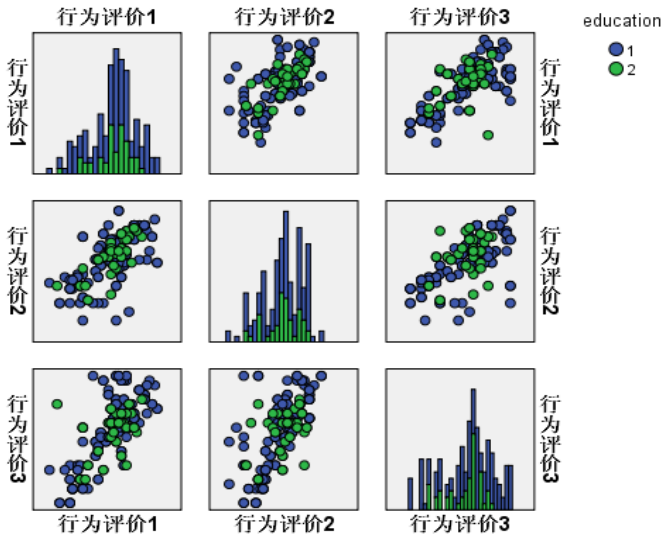
③ 变量为连续变量（积差相关的条件）。

④ 两变量间的关系是线性的。

相关系数与相关程度对应情况表

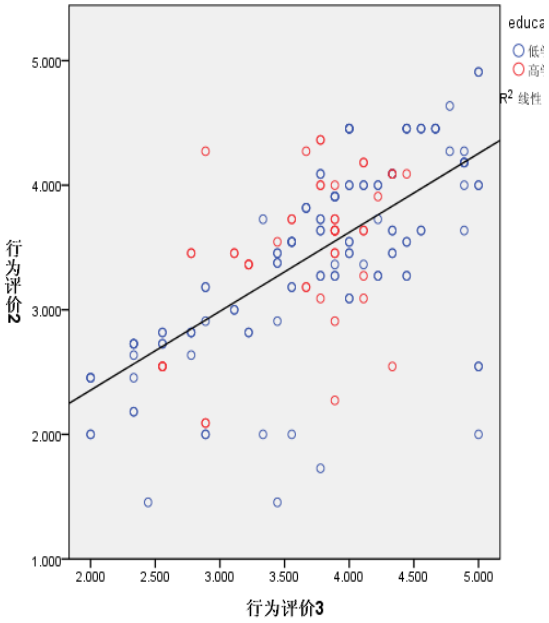
相关系数的绝对值		
0	无相关	
[0　0.3)	弱相关	注意：大样本的情况可能会使弱相关显著。
[0.3　0.5)	低相关	
[0.5　0.8)	显著相关	注意：这里的“显著”有别于统计中的显著。
[0.8　1)	高相关	复杂模型的基础
1	完全相关	系数超过0.96，需注意后续模型变量间的共线性问题。

注：该对应关系需视学科背景的具体情况而定。

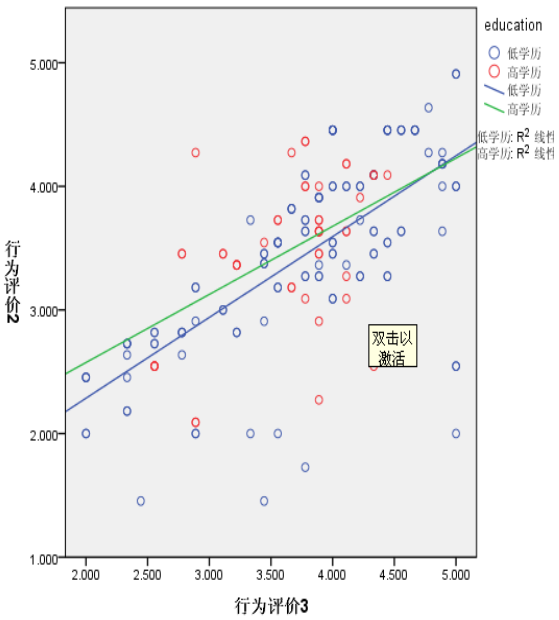


矩阵散点图

散点图提供如下特征：散点的密集程度，反应相关性的大小；散点是否具有线性关系，或线性趋势，还是其他形式，如果是其他形式是否可以转换成线性形式；线性关系之外是否存在异常值及其存在与线性趋势的哪个方向；数据是否存在稀疏问题。



总拟合



子拟合

相关分析：Pearson：适用于两列连续变量的情况。Kendall 的 tau-b：适用于两列有序分类资料。Spearman：是按秩次大小计算的线性相关分析。适用的范围很广，但统计效能要低于 Pearson。

注：其他用于度量有序分类或无序分类的相关分析指标，在分析→描述统计→交叉表→（选项）统计量。

Fisher Z 变换：

若相关系数$r_{yx} = 0.867$，相关系数的抽样分布严格意义上说是正态分布的，但如果相关系数取值偏大，抽样分布通常是负偏分布的，这样导致相关系数的检验出现偏差。

$fisher$ Z=1/2 * ln((1+r_{yx}) / (1-r_{yx}))。

标准误：$\sigma_e = 1/\sqrt{(n-3)}$

$\Rightarrow Z = (z_r - u)/\sigma_e$

注：$r_{yx} = (e^{2z_r} - 1)/(e^{2z_r} + 1)$

这样 Z 和r_{yx}的置信区间都可以计算出来。

（2）偏相关分析

该功能用于解决什么问题：

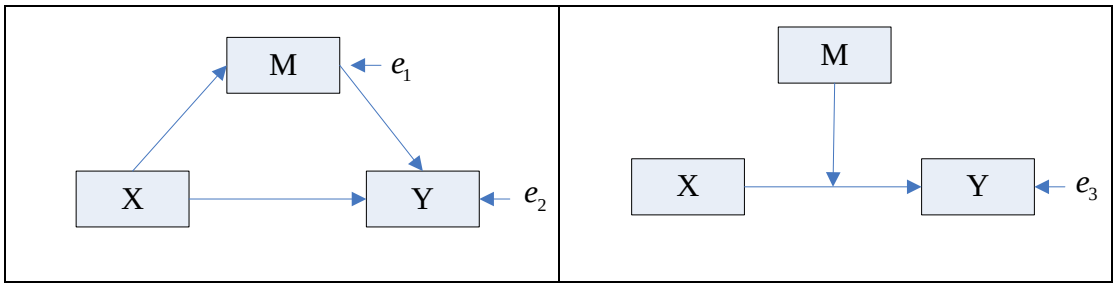
相关系数表达的统计效应中，有可能是来自于其他变量，例如冰棍销售量与性犯罪有显著的相关性，但它们间的关系多半与温度有关，所以如果不控制温度这个因素，就难以探究

冰棍销售量与性犯罪间的真实关系，控制温度后，两者间可能就不显著了。

相关分析只是考虑了两个变量间的关系，偏相关可以同时考虑更多的控制因素，即可以消除其他关联性因素的影响后，在分析两个变量间的关系。一般用于回归分析（中介、调节）的预分析。

偏相关：又称净相关，表示排除其他变量的影响后，计算两个连续变量间的相关。

表现形式：多种表现形式。其他变量可以是中介变量，也可以是调节变量等。



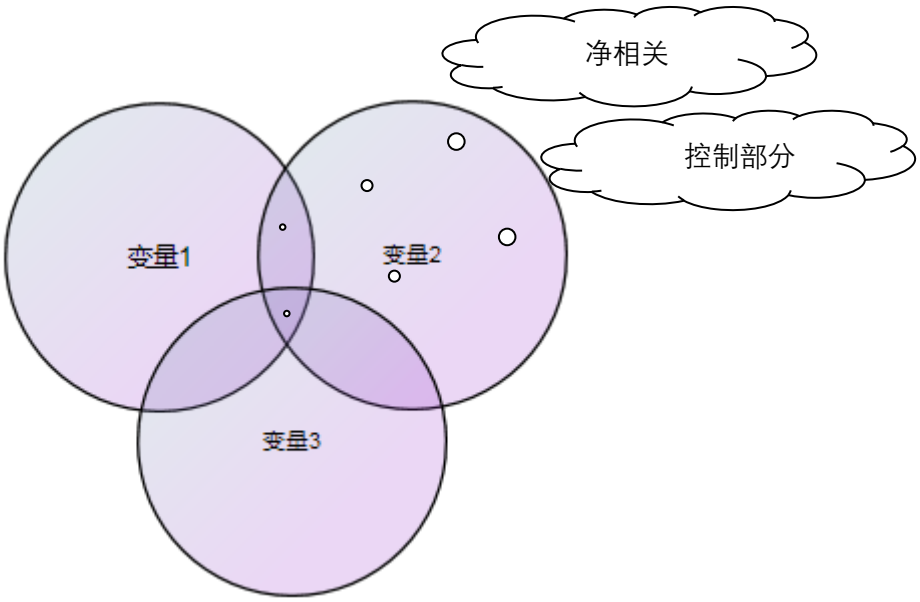
公式：

$$r_{12(3)}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

注：$r_{12(3)}$ 表示在控制比变量 3 的情况下，研究变量 1 与变量 2 间的偏相关系数。

r_{12} 表示变量 1 与变量 2 间的零阶相关系数，其他类似指标以此类推；

实际上就是控制变量不能解释的部分间的零阶相关，如下图：



（3）交互效应模型

H0：简化模型是充分的，H1：全模型是充分的，

简化模型与全模型是嵌套模型。

$$F = ((SSE_{(简)} - SSE_{(全)})/(p+1-k))/(SSE_{(全)}/(n-p-1))$$

注：p是全模型待估参数，k是简化模型待估参数。

模型摘要[c]

模型	R	R 平方	调整后的 R 平方	标准估算的错误	更改统计量				
					R 方变化	F 更改	df1	df2	显著性 F 更改
1	.538[a]	.289	.282	4.27232	.289	39.831	1	98	.000
2	.601[b]	.362	.348	4.06919	.073	11.028	1	97	.001

a. 预测变量：（常量），情绪总分

b. 预测变量：（常量），情绪总分, 适应总分

c. 因变量：绩效总分

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	727.024	1	727.024	39.831	.000[b]
	残差	1788.766	98	18.253		
	总计	2515.790	99			
2	回归	909.635	2	454.817	27.468	.000[c]
	残差	1606.155	97	16.558		
	总计	2515.790	99			

a. 因变量：绩效总分

b. 预测变量：（常量），情绪总分

c. 预测变量：（常量），情绪总分, 适应总分

==========

2.3 线性回归与方

单因素方差分析（单因素 ANONA）

是指将所获得的数据按某些项目分类后，再分析各组数据之间有无差异的方法。即，变异分解过程。

该功能用于解决什么问题:

用于检验单因素不同水平（>=3）时，某因变量均值是否有显著变化的情况，还可以进一步处理不同水平间的精细比较。例如不同受教育水平（假如水平数>=3）的员工绩效是否有差异，及其每种受教育水平两两间的比较。

（1）方差分析原理

方差分析依据的基本原理是方差的可加性原则。

计算 F 统计量过程

①变异分解:

$$SS_T = SS_B + SS_W,\ df_B = k-1, df_W = k(n-1)$$

注:

$$SS_T=\sum\sum X^2-\frac{(\sum\sum X)^2}{nk}$$

表示总体平方和,

$$SS_B=\sum\frac{(\sum X^2)}{n}-\frac{(\sum\sum X)^2}{nk}$$

表示组间平方和,

$$SS_W=\sum\sum X^2-\frac{(\sum X)^2}{n}$$

表示组内平方和。

n 表示总观测数，k 表示水平数。

②计算均方

$$MS_B=SS_B/df_B\ ,\ MS_W=SS_W/df_W$$

注：MS_B 表示组间均方。MS_W 表示组内均方。

③计算 F 比值

$$F=\frac{MS_B}{MS_W}$$

单因素方差分析模型:

$$X_{ij}=\mu+\alpha_i+\varepsilon_{ij}$$

注：μ 是总体均值；α_i 是因素不同水平对因变量的附加效应，其和为 0;

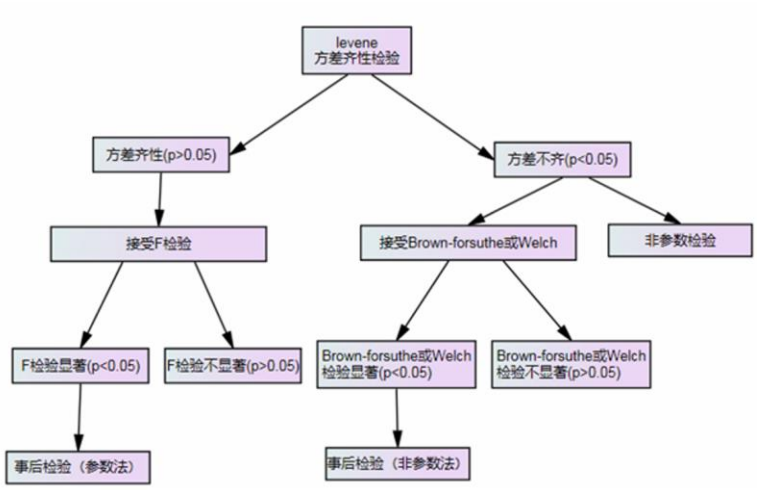
$\varepsilon_{ij}\sim N(0,\sigma^2)$

适用条件：因素水平间的因变量要服从正态分布;

适用于分类水平为两个以上的分类变量;

总体方差相等。

(2) 方差分析步骤



==========

2.4 数据分析流程

(1) 散点图——主体模式：相关、趋势、异常
(2) 相关分析——回归风向标
(3) 回归——系数与 r 方
(4) 残差分析：正态、异常、异方差
(5) 应用：结构与预测。