CDA 统计课程

==========

3.多元线性回归

3.1 多元线性回归的假设

3.2 正态分布问题

3.3 异方差问题与处理

3.4 异常值问题与处理

3.5 共线性问题与处理

3.6 内生性问题与处理

3.7 reg 与 dmreg

==========

3.1 多元线性回归的假设

线性回归模型的假设：

A0:线性：因变量与自变量间的线性关系。

A1: 正交假定：误差项与自变量不相关，其期望为 0，保证无偏性。

*--正态性与异方差在构建模型中的作用较小，对推论影响大---.

A2: 独立同分布：残差间相互独立，且遵循同一分布（iid），要求方差齐性，保证有效性。

*--可以通过残差的独立性检验 Durbin-Watson---.

A3:正态性：因变量的正态性（残差服从正态性）。包括以上假设保证了最佳线性无偏估计。

*---使用残差图诊断---。

推理统计：对参数建立假设，使用假设检验校验，即点估计（平均数或系数）、点估计的准确性（标准误）、使用点估计和标准误构建统计量（t、F 等），做统计推断。

无偏性——如果估计量$\hat{\theta}$的期望，等于被估计的总体参数θ估计量就是无偏的，样本均值$\bar{x}$就是总体均值μ的无偏估计量，而样本调整方差$S^2 = \sum_{i=0}^{n}(x_i - \bar{x})^2/(n-1)$才是整体方差的无偏估计量。

有效性——当一个总体参数存在多个无偏估计量时，还需看所在抽样分布是否具有尽可能小的方差，称之为有效性。

一致性——有些总体未知参数不一定存在无偏估计量，而有些却存在不止一个无偏估计量。当样本容量 n 不断增大，估计量越来越接近总体参数的真实值时，称之为一致性。

==========

3.2 正态分布问题

短期绩效中，变量绩效有 100 个样本点的数据，并计算均值和标准差两个参数，使用该参数模拟正态分布。这样我们就得到图 2-8 左侧的图形，分别将原始数据和模拟的正态分布数据绘制直方图。模拟数据是正态分布随机数，理论上可以视其为正态分布的“标准”，这样原始数据与“标准”间的偏差，就构成了判断是不是正态分布的依据，因此两组数据相减产生右侧偏差 d 的直方图。讲这种数据变量四维滴。正态分布检验就是来检验最大 d 是否等于 0，这构成了假设检验中的原假设。读者可以使用 K-S 检验或 S-W 检验等校验其是否符合正态分布（操作：分析→描述统计→探索或分析→非参数检验）。不过就便利性、检验考察的严格度而言，笔者还是推荐使用 PP 图，此处用于描述 PP 图检验原理、解读及使用状况。

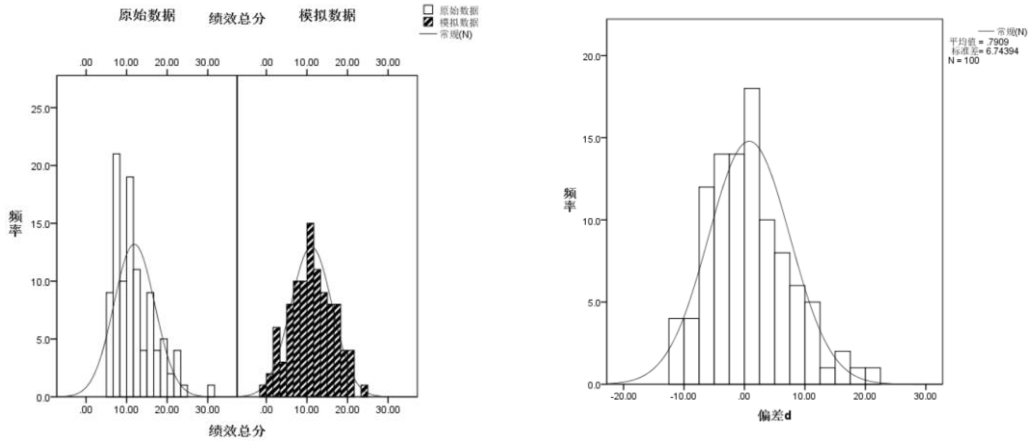


图 2-8 正态分布检验

图 2-8 所示的模拟文件中，分别确定原始数据的实际累积概率和模拟数据的理论累积概率，观察图形中的数值的匹配程度，检验分布特征，如图 2-9。图 2-8 的两幅图和图 2-9 的图形性质相同。左侧 PP 图是根据原始数据（散点）和模拟数据（对角线）分别绘制的散点图，右侧同样也是原始数据与模拟数据对应的点相减产生的偏差图。PP 图的读法是，原点是否紧密缠绕或聚集在理论线上，若偏离的幅度很小，表示原始数据符合正态分布，反之不符合，例如第 1 行观测值相比第 23 行更容易导致分布假设问题。但问题是这个幅度多小为小，很遗憾 PP 图没有提供具体的量化指标，但是 PP 图的偏差图（去势图）提供了经验界值。

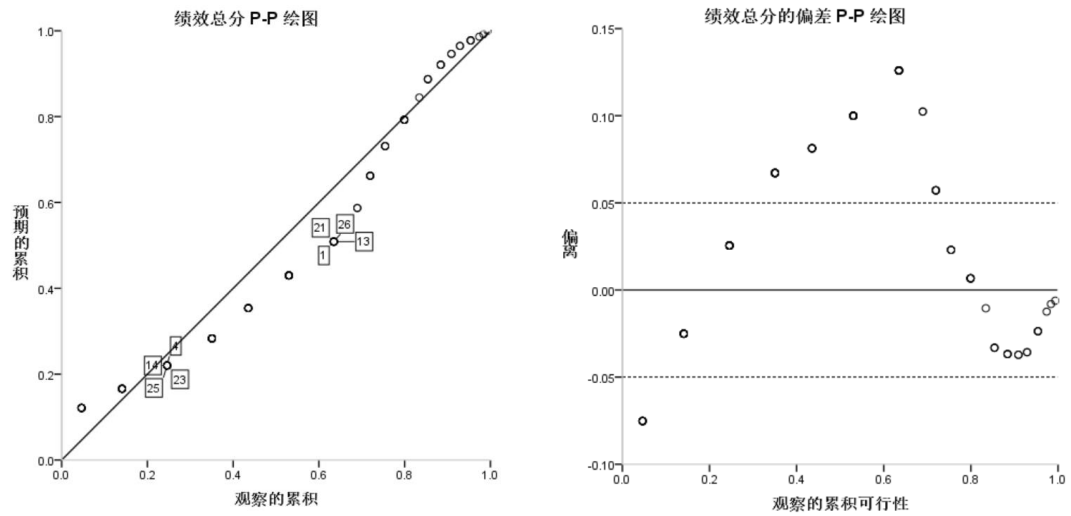


图 2-9 PP 图的分布检验

右侧偏差图，散点表示实际原始值与理论值之差。在正态分布情况下，散点应该围绕 0 点上下小幅的均匀分布，差值波动控制在[-0.05 0.05]间为宜[1].

==========

3.3 异方差问题与处理

方差不齐—加权最小二乘法

*-- log 变换、加权最小二乘法、稳健回归、---.

- 变量变换的方法，如 sqrt（泊松分布或罕见的计数问题）、log（方差大于均值）等变换，更一般的情况见 box-cox 变换。
- 权重估计：$w_i = 1/\sigma_i^2$, 且 $\sigma_i^2 = kx_i^2$ 因为观测值权重是观测值的误差方差的倒数，并且观测值的误差方差随预测变量的变化而发生系统变化，因此权重与预测变量的更一般的形式可以表示为 $w_i = 1/x_i^m$，故当下需估计 m 的最优取值。

第一步：输出 abs（残差）与所有自变量的等级相关系数，选择最大的相关变量作为权重变量 X^m。

第二步: 选择 auxiliary_materials1 最为权重变量。并产生的最佳权重值保存为新变量。如 WGT_1。并将 WGT_1 作为 OLS 估计的加权变量。

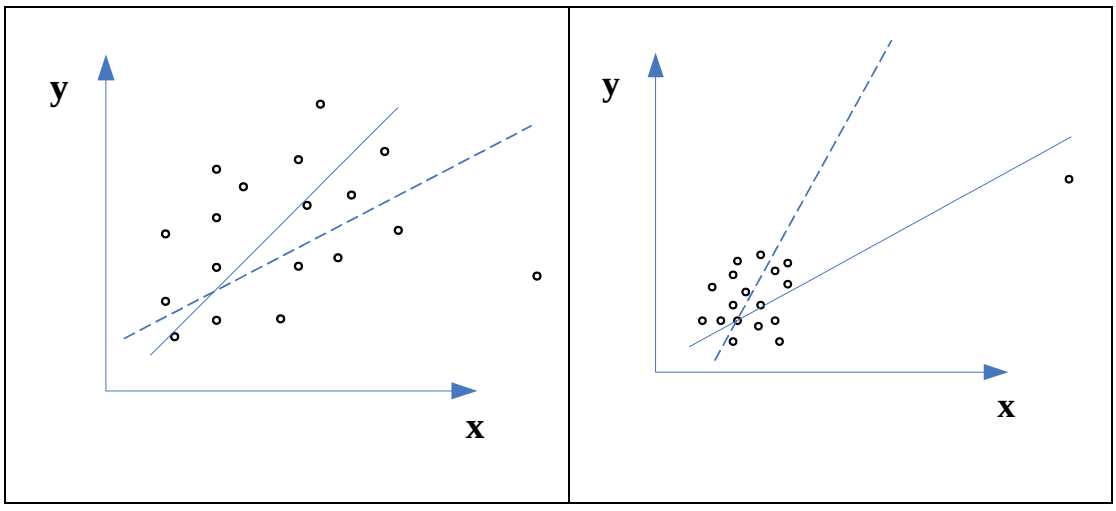
软件自动根据对数似然值判断最优的模型。

第三步：变成预测值与残差绘制散点图。

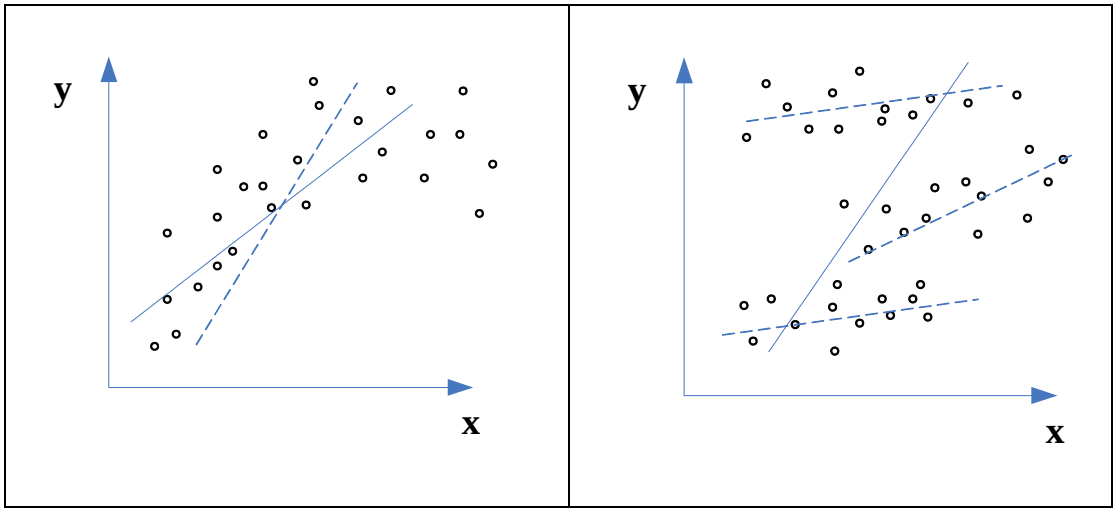
==========

3.4 异常值问题与处理

----------以下几种散点图的对比----------



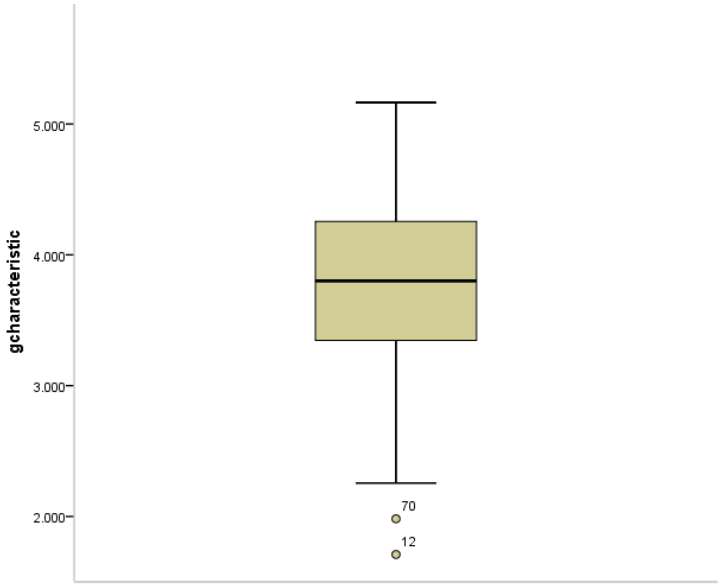
[1] 参见《SPSS 统计分析基础教程》的数据的图形展示。



---

a、单变量验证

箱体图



统计量

gcharacteristic

N	有效	182
	缺失	0
均值		3.71696
中值		3.80024
众数		4.509
标准差		71 971
方差		.517

箱体图说明:

①箱图中间的粗线表示中位数。

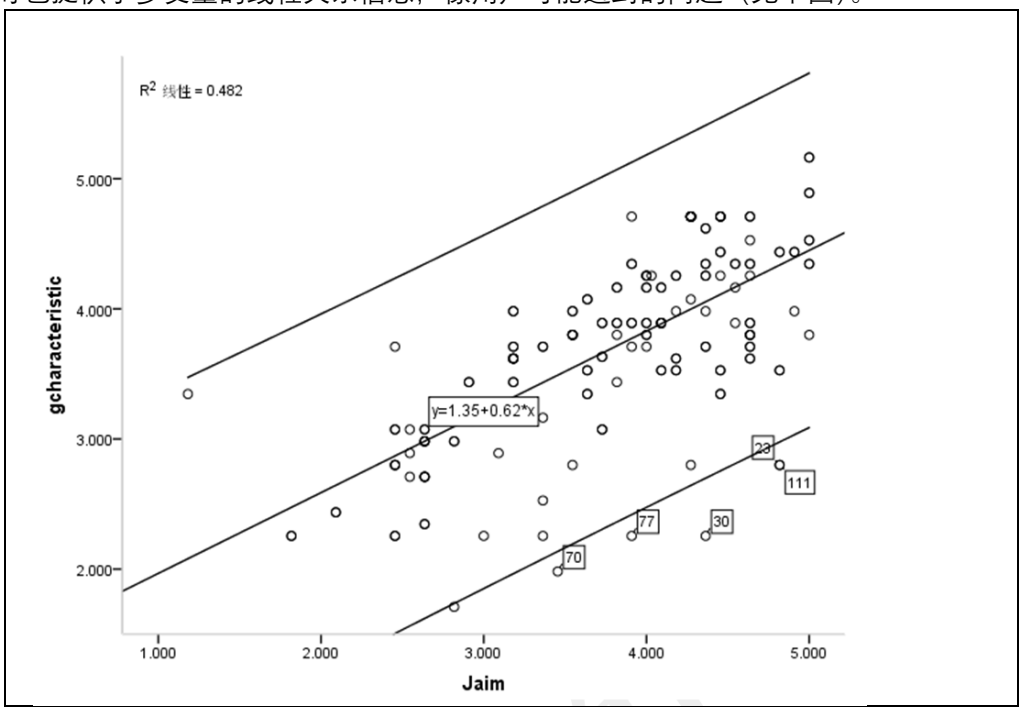
②箱体（灰绿色）的上、下线表示四分位数，即 25%和 75%的百分位数，可以看出四分位距可以反映整体样本中间一半的数据分布情况。

③箱体外的上下线表示除去异常值外的最大值和最小值。

④与箱体（灰绿色）的上、下线的距离超过四分位距的 1.5 倍即被视为异常值（用圆圈表示），超过 3 倍的即被视为极端值（用星号表示）。

b、双变量：散点图

散点图用于寻找数据的主体模式，其个体的 99%的置信区间，可以帮助用户找到可能的异常信息。实际上，常见的多变量模式分析中多半也借助散点图功能，来检查异常信息，当然同时也提供了多变量的线性关系信息，像用户可能遇到的问题（见下图）。



下面根据因变量与自变量数量、测量标准的综合信息，来提取主成分或某种得分（本质上都是提取主成分），作为绘制散点图的依据。下图以“因变量个数与类型“为核心。

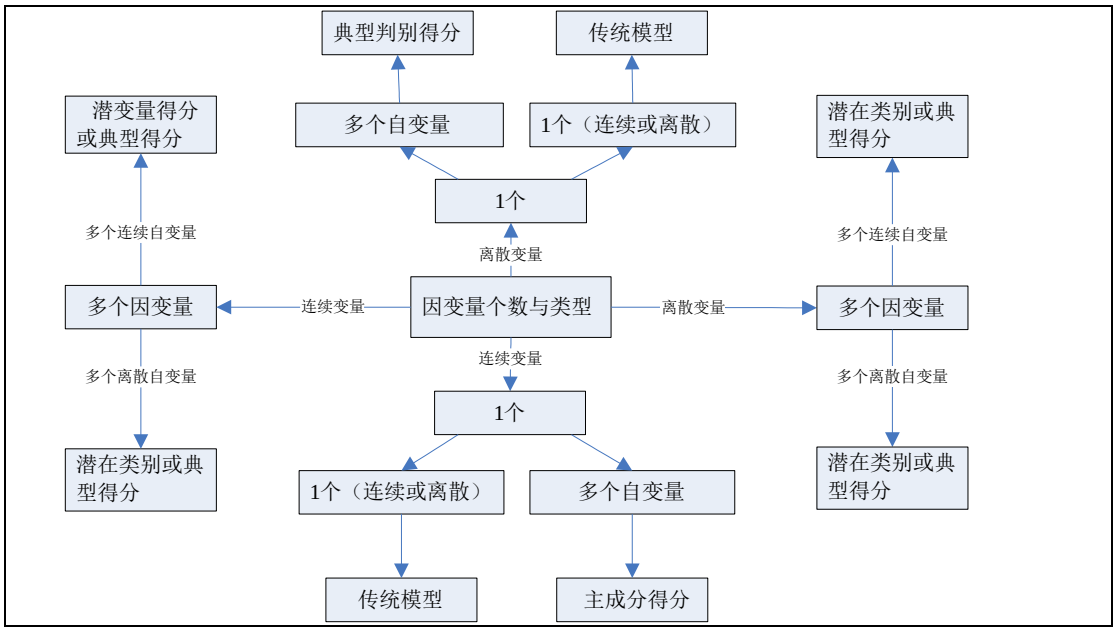
①12 点钟方向：因变量只有一个，且为离散性变量，多个自变量使用典型判别分析，那么因变量（转化成虚拟变量）、自变量位置上均可以提取典型得分，这些得分用于绘制二维的散点图。当然如果是一个自变量，这自然好处理（下面也不在强调一个自变量的情况）。

②6 点钟方向：因变量只有一个，且为连续性变量，多个自变量使用主成分分析，对多变量提取主成分信息，或使用总因子得分，这些降维后的信息就可以用于绘制二维的散点图了。

③3 点钟方向：因变量有多个，且为离散性变量这的种情况，多数要视具体情况而定，但一般性的方法有助于启示用户方法的选择，如果于此相对的有多个自变量可以使用的方法，如潜在类别分析技术的处理、典型得分（也可以使用典型判别），有时根据变量及其类别间的线性关系等信息也可以使用最优尺度转换，从而寻求比较简单的主成分提取等方法，那么最终就可以获得二维关系的散点图。

④9 点钟方向：因变量有多个，且为连续性变量，于此相对的也有多个自变量，常用的处理过程，如使用结构方程的潜变量得分、典型得分在变量间的双向关系和单向关系时均适用，主成分分析等。

由于多维信息在理解上的限制，总是寻求变量间的低维关系来绘制图形，以期帮助分析人员了解变量间的关系，预分析中介于所见关系的问题，其实大多都在寻求此道。



注：总的来说，就是实现多变量的降维（主成分）。

c、多变量：标识异常个案：聚类分析

输出结果：

异常个案原因列表

原因:1

案例	原因变量	变量影响	变量值	变量范数
51	适应总分	.897	1.00	8.6667
15	适应总分	.568	22.00	12.7419
58	绩效总分	.713	17.00	9.5362

异常索引：该值越大表示越可能为异常值。

变量影响：该个案的影响程度指标。

变量值：就是原始值。

变量范数：可以理解为该个案对应的对等组中的集中趋势

----------异常值处理----------

异常值处理的常用方法：

（1）直接将该条观测删除

在 SPSS 软件里有 2 种不同的删除方法，整条删除和成对删除。

当然，这种方法简单易行，但缺点也很明显，首先我们经常会遇到的情况是观测值很少，这种删除会造成样本量不足，其次，直接删除的观测很多，也可能会改变变量的原有分布，从而造成统计模型不够稳定。

（2）暂且保留，待结合整体模型综合分析

通常我们观测到的异常值，有时在对于整个模型而言，其异常性质并没有观测到的明显，因此最好综合分析一下，像回归分析，我们经常利用残差分布信息来判断模型优劣，

残差有没有超出经验范围（+3 标准差），呈现什么分布等，另外对于整个模型而言，会有一些指标像 Mahalanobis、Cook's、协方差比率等可以提供某条观测或整体的拟合信息，这些指标也会提示分析人员的异常值信息。如果对于整个模型而言，并不是很明显时，建议保留。

（3）如果样本量很小，可以考虑使用均值或其他统计量取代

这不失为一种折中的方法，大部分的参数方法是针对均值来建模的，用均值取代，实际上克服了丢失样本的缺陷，但却丢失了样本“特色”，可以说是不大不小的错误。当然如果是时序数据，用于取代的统计量，可供选择的范围就会多一些，可以针对序列选择合适的统计量取代异常值，也较少存在上述问题。

（4）将其视为缺失值，利用统计模型填补

该方法的好处是可以利用现有变量的信息，对异常值（缺失值）填补。不过这里最好要视该异常值（缺失值）的特点而定，例如需视是完全随机缺失、随机缺失还是非随机缺失的不同情况而定。

（5）不做过多处理，根据其性质特点，使用稳健模型加以处理

如果按参数性质分的话，可以将稳健方法分为参数、非参和半参 3 种情况，这总体上与参数的假设、优点一样。

（6）使用抽样技术或模拟技术，接受更合理的标准误等信息

抽样样本（SPSS 默认是 1000）所计算出的均值的标准误，一般来说会更合理，这可以有效应对异常值的影响，但前提是原始样本量不能太少（小于 10），小样本的结果不够稳定。另外模拟技术可以利用先验分布特征和样本信息来构建事后预测的概率分布，进行事后模拟，这种技术现在发展的很好，在异常值的应对中，表现良好。

==========

3.5 共线性问题与处理

*---处理方法：项目合并、主成分回归、逐步回归、偏最小二乘、分位数回归等----.

----------岭回归相关信息----------

岭回归是最小二乘法的一种（k=0）,属于有偏估计的方法，主要是以损失少部分信息和精度的前提下，更符合实际情况的回归方程。

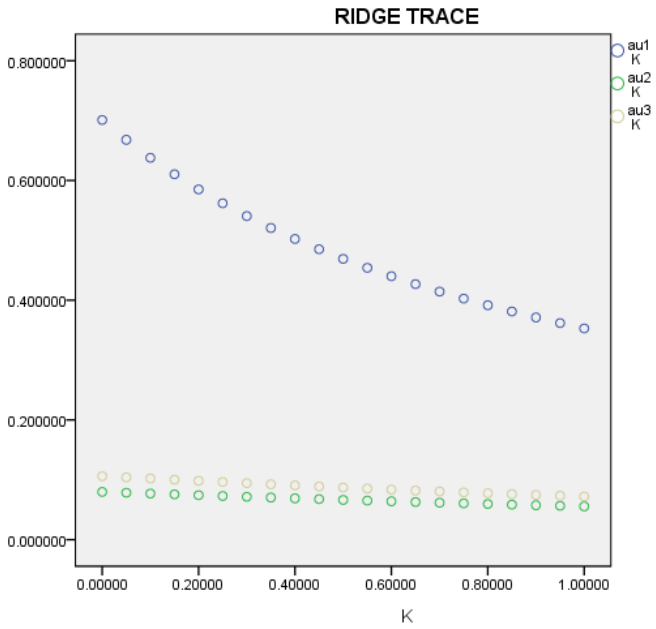
当自变量存在严重共线性时，并不会导致矩阵的行列式等于零，但会非常接近于零。这种情况称为非正定或奇异，这时的 X'X 就是奇异的。但如果将 X'X 加上正常数矩阵，奇异性就会得到有效改善，K 就是该系数的估计值。

文件：商品材质 1.sav

岭回归程序:

R-SQUARE AND BETA COEFFICIENTS FOR ESTIMATED VALUES OF K

K	RSQ	au1	au2	au3
.00000	.53441	.701120	.079777	.105964
.05000	.53328	.667971	.078471	.104118
.10000	.53031	.637844	.077128	.102202
.15000	.52596	.610340	.075766	.100251
.20000	.52060	.585129	.074397	.098290
.25000	.51450	.561934	.073034	.096340
.30000	.50786	.540521	.071684	.094413
.35000	.50086	.520691	.070352	.092519
.40000	.49360	.502274	.069044	.090665
.45000	.48620	.485123	.067762	.088854
.50000	.47872	.469112	.066508	.087090
.55000	.47123	.454129	.065285	.085375
.60000	.46376	.440079	.064092	.083709
.65000	.45635	.426877	.062931	.082091
.70000	.44903	.414447	.061802	.080523
.75000	.44181	.402724	.060703	.079003
.80000	.43472	.391649	.059636	.077530
.85000	.42776	.381169	.058598	.076103
.90000	.42094	.371238	.057591	.074721
.95000	.41426	.361813	.056613	.073382
1.0000	.40773	.352856	.055662	.072084



岭迹图

解释：根据 K 需最小、R2 需最大的原则选择最优值。

岭回归一般主要应用于：

1）帮助用户判断使用传统的 OLS 估计是否合适，例如，如果岭迹图显示大多预测变量的变化幅度较大，预示普通 OLS 估计会很差。

2）筛选自变量，如果岭迹图显示自变量的绝对值系数较小或波动不稳定，可以考虑删除。

3）对比普通 OLS 与岭回归分析的区别，了解共线性的影响大小及其方向问题。

==========

3.6 内生性问题与处理

==========

3.7 reg 与 dmreg

速度与精确度的关系

Reg——精确度

Dmreg——速度