第一章　回归分析简介

1. 1 曲线拟合

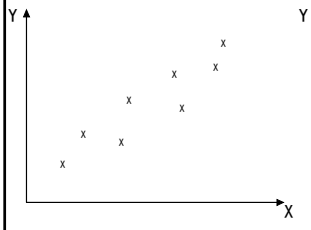
1.1.1 散点图

设有 n对观测值$(X_i,Y_i)\ i=1,2,\cdots n.$“确定”两个变量 X,Y 之间所存在的关系。

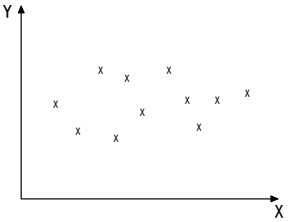
通常作法：

1.作散点图；

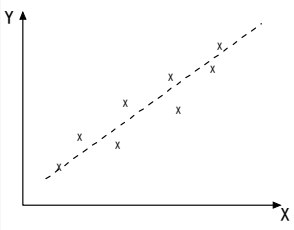
2.根据散点图尽量拟合一条“优美”曲线，使这些点尽可能“趋近”这条曲线。



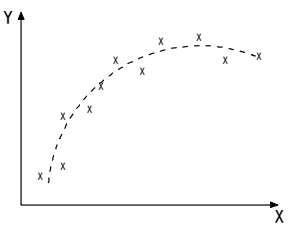
X,Y相关的一种散点图



X,Y不相关的一种散点图



X,Y呈线性关系



X,Y呈非线性关系

1.1.2 拟合曲线的选择

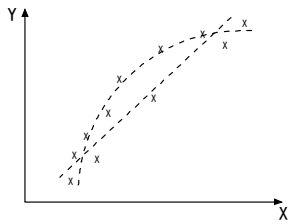
- 拟合一条“优美”的曲线通常指拟合一条光滑的曲线；
- 由于许多不可控因素使变量发生随机波动，不要期望点与曲线完全拟合；
- 即使两变量之间存在确定关系，由于测量误差，这种偶然性的波动仍然会在散点图上表现出来。

- 按经验规律或有关理论等选择拟合曲线的类型

例 1：设一组试验数据(X,Y)，其中X是通过某电阻的电流安培数，Y是加在该电阻的电压伏特数。由欧姆定律X,Y有关系$Y=r\cdot X$，过原点呈直线型的散点图可印证这个定律。用过原点的直线去拟合这组数据，还可以用直线的斜率去估计电阻 r的值。

例 2：对保持恒温的一定质量的气体得到一组试验数据(V,P)，根据物理学原理V,P有关系$PV^r=c$，取对数得到线性关系$\log P=\log c-r\log V$，因此$\log c$和 r可以由拟合试验数据$(\log V,\log P)$的直线估计出来。

- 当没有经验规律和有关理论帮助时，确定应该拟合曲线的类型有时是困难的。



拟合同一组数据的两条不同曲线

一般说来若假设拟合的曲线形状已知，只是包含若干未知参数，由已知数据来估计这些未知参数，此方法称为参数的方法；若对拟合的曲线形状没有假定，一般用非参数的方法来拟合曲线。

1.2 回归分析

1.2.1 协变量与响应变量

在现实世界中存在大量这样的情况：两个或多个变量之间有一些联系，但可能没有确切到可以严格决定的程度。

例 1：人的身高X与体重Y有联系，一般表现为X大时Y也倾向于大，但由X不能严格决定Y。

例 2：一种农作物的亩产量Y与其播种量X_1施肥量X_2有联系，但X_1, X_2不能严格决定Y。

以上例子以及类似的例子中，Y通常称为响应变量(response)(或因变量、预报变量)，X, X_1, X_2等则称为协变量(covariates)(或自变量、预报因子)。协变量可以是随机的，也可以是非随机的，通常是可以观测的。

1.2.2 回归函数

现设在一个问题中有响应变量Y，协变量$X_1, \cdots X_p$。可以设想响应Y由两部分构成：一部分由$X_1, \cdots X_p$的影响所致，这一部分可以表为$X_1, \cdots X_p$的函数形式$f(X_1, \cdots X_p)$；另一部分则由其他众多未加考虑的因素，包括随机因素的影响所致，这部分视为一种随机误差，记为e。

于是得到模型：

$$Y = f(X_1, \cdots X_p) + e$$

这里e作为随机误差要求$Ee = 0$。

上式也可写为

$$f(X_1, \cdots X_p) = E\left(Y \middle| X_1, \cdots X_p\right)$$

称为Y对$X_1, \cdots X_p$的理论回归函数(regression function)。回归函数前面所加“理论”两字是为区分由数据估计所得的回归函数(称为经验回归函数)。

在实际问题中，理论回归函数一般总是未知的，统计回归分析的任务在于根据 $X_1,\cdots X_p$ 和 Y 的观测值去估计回归函数及讨论与此有关的一些统计推断问题。所用的方法在很大程度上取决于对模型中回归函数 f 及随机误差 e 所作的假定。若对回归函数 f 的数学形式并无特殊假定，称为非参数回归(non-parametric regression)；若假定 f 的形式已知，只是其中若干参数未知，这种情况称为参数回归(parametric regression)。

一般说来 在参数回归中，若 f 关于未知参数是线性的，称为线性回归(linear regression)；若关于参数是非线性的，称为非线性回归(nonlinear regression)。

对于随机误差 e，已经假定其均值 $Ee=0$，其方差 $Var(e)=\sigma^2$ 是模型的一重要参数。由于

$$E\left(Y-f(X_1,\cdots X_p)\right)^2=Ee^2=\sigma^2,$$

因此 σ^2 越小，用回归函数 $f(X_1,\cdots X_p)$ 逼近 Y 的均方误差(mean square error)就越小。

误差方差 σ^2 的大小主要由以下两点决定：

1).选择协变量时，是否把对响应变量有重要影响的那些因素都包括了；

2).回归函数的形状是否选择准确。

1.3 线性回归模型与最小二乘法

1.3.1 线性回归模型(linear regression models)

设响应变量 y 与协变量 $x_1,\cdots x_p$ 有关系

$$y=x_1\beta_1+\cdots x_p\beta_p+e, Ee=0,$$

现有 n 次观测，即

$$y_i=x_{i1}\beta_1+\cdots x_{ip}\beta_p+e_i, i=1,\cdots,n.$$

这里 $(y_i,x_{i1},\cdots x_{ip})$ 为已知，$(\beta_1,\cdots,\beta_p)$ 为未知非随机的参数，e_i 是第 i 次观测的随机误差，且假定 $Ee_i=0$，$Cov(e_i,e_j)=\sigma_{ij},1\le i,j\le n.$

通常写成简洁的矩阵形式。令

$$X_{n\times p}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$Y=(y_1,\cdots,y_n)',\beta=(\beta_1,\cdots,\beta_p)'$$

$$e=(e_1,\cdots,e_n)',\Sigma=Cov(e)=(\sigma_{ij})_{n\times n}.$$

这样得到矩阵表示：

$$\begin{cases} Y=X\beta+e, \\ Ee=0,Cov(e)=\Sigma. \end{cases} \quad (1)$$

模型(1)描述变量之间线性相关关系的一类统计模型，称为线性回归模型(注意关于未知参数β是线性的)。向量 Y通常称为观测向量，矩阵X通常称为设计矩阵(design matrix)。

设计矩阵X的秩$rank(X)$可以等于p(此时X称为满秩的)，也可以小于 p。一般线性回归模型设计矩阵X是满秩的，方差分析模型和协方差分析模型的设计矩阵X不是满秩的。设计矩阵 X是否为满秩，对模型(1)的参数估计问题会产生很大的影响。

1.3.2 GM 条件

对于误差协方差阵$\Sigma=Cov(e)$，若为未知，则难以由 n次观测数据去估计模型(1)的 p个未知系数$\beta_1,\cdots,\beta_p$以及$n(n+1)/2$个未知量$\sigma_{ij},1\le j\le i\le n$。故还要对误差协方差阵$\Sigma$作些假设，例如假设 Σ已知或者$\Sigma=\sigma^2\Sigma_0$，而 Σ_0已知等。

通常我们认为 n次观测的误差是等方差的，且互不相关，即对误差协方差阵假设$\Sigma=\sigma^2 I_n$(其中I_n表示 n阶单位阵)，此假设条件称为 Gauss-Markov 条件，简写为 GM 条件。此时模型未知参数$p+1$个，一般说来，只要观测值多于未知参数个数就可以对这些未知参数作出估计。

1.3.3 最小二乘法(least squared method)

考虑如下线性模型：

$$\begin{cases} Y = X\beta + e, \\ Ee = 0, Cov(e) = \sigma^2 I_n. \end{cases}$$

估计未知系数β的基本出发点是：参数的真值应该使模型误差$e = Y - X\beta$达到最小。令$Q(\beta) = \|e\|^2 = \|Y - X\beta\|^2$来度量模型误差的大小，则$\beta$的估计应最小化$Q(\beta)$，即

$$\hat{\beta} = Arg\min_{\beta} \|Y - X\beta\|^2 。$$

上述确定未知参数 β的方法称为最小二乘法，所得到的估计称为最小二乘估计(least squared estimate)，简写成 LSE。

例 1：回忆前面提到的曲线拟合问题，有 n个点$(x_i, y_i), 1 \le i \le n$，假设拟合的曲线形式已知为$f(x, \beta)$(注意$\beta$是向量)，按照最小二乘法的原理未知参数的真值应该使的被拟合的曲线的纵向偏差的平方最小，即

$$\hat{\beta} = Arg\min_{\beta} \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \quad 。$$

1.4 方差分析

在实际中经常要对在不同条件下进行试验或观察得到的数据进行分析，以判断不同条件对结果(响应变量)有无影响。此时协变量往往表示某种效应(条件)存在(成立)与否，因而往往取 0，1 两个值。在统计学上称这类分析为方差分析(analysis of variance，ANOVA)。

例 1：某农业科研机构欲比较三种小麦品种的优劣，设计了一种比较试验。为保证试验结果的客观性，他们选择了六块面积相等、土质肥沃程度一样的田地。每一种小麦播种在两块田地上，并给予几乎完全相同的田间管理。设用y_{ij}表示第i种小麦的第j块田地的产量，则可以对其作如下的分解：

$$y_{ij}=\mu+\alpha_i+e_{ij},\ i=1,2,3;\ j=1,2$$

其中μ表示总平均，α_i表示采用第i个小麦品种对产量的影响效应，e_{ij}表示其它为控制因素及各种随机误差的效应。

写成矩阵形式：

$$\begin{pmatrix} y_{11}\\ y_{12}\\ y_{21}\\ y_{22}\\ y_{31}\\ y_{32}\end{pmatrix}=\begin{pmatrix} 1&1&0&0\\ 1&1&0&0\\ 1&0&1&0\\ 1&0&1&0\\ 1&0&0&1\\ 1&0&0&1\end{pmatrix}\begin{pmatrix}\mu\\ \alpha_1\\ \alpha_2\\ \alpha_3\end{pmatrix}+\begin{pmatrix} e_{11}\\ e_{12}\\ e_{21}\\ e_{22}\\ e_{31}\\ e_{32}\end{pmatrix}。$$

若记 $Y=(y_{11}\quad y_{12}\quad y_{21}\quad y_{22}\quad y_{31}\quad y_{32})'$，

$\beta=(\mu\quad \alpha_1\quad \alpha_2\quad \alpha_3)'$，

$e=(e_{11}\quad e_{12}\quad e_{21}\quad e_{22}\quad e_{31}\quad e_{32})'$，

$$X=\begin{pmatrix} 1&1&0&0\\ 1&1&0&0\\ 1&0&1&0\\ 1&0&1&0\\ 1&0&0&1\\ 1&0&0&1\end{pmatrix},$$

则可以写成：

$$Y=X\beta+e,$$

回到上节线性回归模型的形式，不过设计矩阵元素有其特点，非 0 即 1。例子所引进的模型是方差分析模型中最简单的一种，称为单因素方差分析，因为只涉及到“小麦品种”这一个因素的影响。在实际中还涉及到两个或多个因素的影响。

例 2：在例 1 中，一般情况下很难找到肥沃程度完全一样的田地。考虑到土质对产量的影响也是不可忽略的。一般，从若干试验田地中选取土质肥沃均匀的b块地(在试验设计中把这种块称为区组,block)，再将每块等分成 3 小块，称为试验单元，在每个试验单元上种植一种小麦。用y_{ij}表示第i种小麦的第j个区组田地的产量，则可以对其作如下的分解：

$$y_{ij}=\mu+\alpha_i+\beta_j+e_{ij},\ i=1,2,3;\ j=1,2,\cdots b$$

其中μ，α_i，e_{ij}的含义与例 1 一样，这里β_j表示第j个区组田地对产量的影响。

这样就得到一个两向分类模型。此模型也可以写成线性回归模型的形式：$Y=X\beta+e$，其中Y,e仿照前面例1的类似记号，$\beta=(\mu\quad \alpha_1\quad \alpha_2\quad \alpha_3\quad \beta_1\quad \cdots\quad \beta_b)'$，设计矩阵$X$为$3b\times(4+b)$矩阵，定义如下

$$X=\begin{pmatrix}1&1&0&0&1&&\\ \vdots&\vdots&\vdots&\vdots&&\ddots&\\ 1&1&0&0&&&1\\ 1&0&1&0&1&&\\ \vdots&\vdots&\vdots&\vdots&&\ddots&\\ 1&0&1&0&&&1\\ 1&0&0&1&1&&\\ \vdots&\vdots&\vdots&\vdots&&\ddots&\\ 1&0&0&1&&&1\end{pmatrix}$$

方差分析模型认为由于各种因素的影响，研究所得的数据呈现波动状。造成波动的原因可分成两类，一是不可控的随机因素，另一是研究中施加的对结果形成影响的可控因素，例如例1中的“小麦品种”。在方差分析中，这些因素也称为因子(factor)，其影响称为效应(effect)，因子的不同状态称为水平(level)。

方差分析的基本思想是：通过分析研究不同来源的变异(方差)对总变异(方差)的贡献大小，从而确定可控因素对研究结果影响力的大小。其目的是通过数据分析找出对该事物结果有显著影响的因素，各因素之间的交互作用，显著影响因素的最佳水平等。

1.5 协方差分析

当知道有些协变量会影响响应变量，却不能够控制或不感兴趣时，可以在实验处理前予以观测，然后在排除这些协变量对观测变量影响的条件下，分析可控变量因素对观测变量的作用，从而更加准确地对控制因素进行评价。此方法称为协方差分析(analysis of covariance，ANCOVA)。

例1：为研究三种饲料对猪的催肥效果，用每种饲料喂8头猪一段时间，测得每头猪的初始重量和增重(如下表)，现分析三种饲料对猪的催肥效果是否相同。

由于饲料是可以控制的，但猪的初始重量是人为无法控制的，因此要把猪的初始重量作为协变量考虑进来，进行协方差分析。令y_{ij}表示第i种饲料后第j头猪增重的体重，x_{ij}表示其初始体重，μ表示总平均，α_i表示采用第i种饲料对猪催肥的影响效应，e_{ij}表示其它为控制因素及各种随机误差的效应。

	初始体重	增重	饲料类型
1	15	85	1
2	13	83	1
3	11	65	1
4	12	76	1
5	12	80	1
6	16	91	1
7	14	84	1
8	17	98	1
9	17	97	2
10	16	90	2
11	18	100	2
12	18	95	2
13	21	103	2
14	22	106	2
15	19	99	2
16	18	94	2
17	22	89	3
18	24	91	3
19	20	83	3
20	23	95	3
21	25	100	3
22	27	102	3
23	30	105	3
24	32	110	3

可以建立如下模型：

$y_{ij}=\mu+\gamma x_{ij}+\alpha_i+e_{ij}$，$i=1,2,3;\ j=1,2\cdots 8$。

令

$$Y=\left(y_{11}\ \cdots\ y_{18}\ \ y_{21}\ \cdots\ y_{28}\ \ y_{31}\ \cdots\ y_{38}\right)',$$

$$\beta=\left(\mu\ \ \gamma\ \ \alpha_1\ \ \alpha_2\ \ \alpha_3\right)',$$

$$e=\left(e_{11}\ \cdots\ e_{18}\ \ e_{21}\ \cdots\ e_{28}\ \ e_{31}\ \cdots\ e_{38}\right)',$$

则写成矩阵形式为：$Y=X\beta+e$，回到线性回归模型的形式，其中设计矩阵X为24×5的矩阵，定义为

$$X=\begin{pmatrix}1 & x_{11} & 1 & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{18} & 1 & 0 & 0\\ 1 & x_{21} & 0 & 1 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{28} & 0 & 1 & 0\\ 1 & x_{31} & 0 & 0 & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{38} & 0 & 0 & 1\end{pmatrix}$$

1.6 线性混合效应模型

为引入此模型，先从一个例子着手。

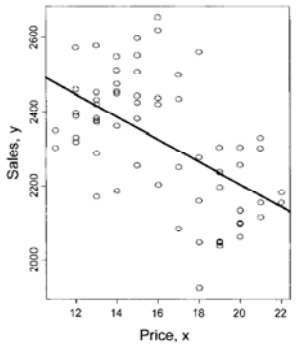
例 1：(销售量-价格模型)

为研究销售量与价格之间关系，选择K种商品，收集到这些商品的销售量与价格的数据$(y_i,x_i)_{i=1}^n$。按照传统线性回归模型，认为商品之间是没有差异的，n次观测是独立的，建立线性回归模型：

$$y_i=\alpha+\beta x_i+e_i,\quad 1\le i\le n$$

这里假设e_i *i.i.d*为零均值，共同方差为σ_e^2。

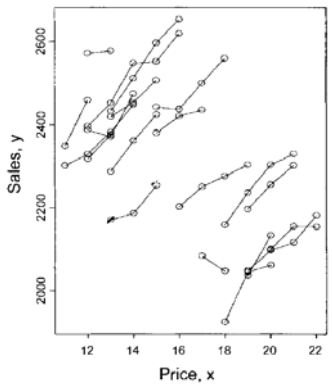
按照此模型，拟合回归直线，其斜率为负。



但事实上不同商品之间存在很大差异。如果按照商品种类将数据分组，记为(y_{ij}, x_{ij})，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，其中$\sum_{i=1}^{K} n_i = n$。对每组中的商品，如果分别建立其销售量与价格的线性回归模型，即假定每组商品有自己特定的模型

$y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，

这里假设e_{ij} *i.i.d* 为零均值，共同方差为σ_e^2。将同组数据点连接的散点图(如下)，其显示大都斜率是正的。



相当于分组去拟合，得到不同组的商品各自的销售量与价格之间的关系。这似乎与最初的目标不一致。最初目标是想宏观的来看商品销售量与价格的关系，而不是特定某类商品销售量与价格的关系。

这样，如果假定每组中的价格对销售量的影响是一样的，不同在于各组截距不一样，得到如下模型：

$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，

这里假设e_{ij} *i.i.d* 为零均值，共同方差为σ_e^2。这样做法并没有解决上面所说的问题，不同的截距项$\{\alpha_i\}_{i=1}^{K}$也只反应了这K组不同商品之间的差异。并不代表总体商品之间的差异。注意如果$\alpha_i \equiv \alpha$，又回到最初的线性回归模型。

一种观点是：将这K组商品看成总体商品中随机抽取的K组商品，$\{\alpha_i\}_{i=1}^{K}$代表每组商品之间的差异。

为使模型简单，假设$\{\alpha_i\}_{i=1}^K$ *i.i.d*均值为α，方差为σ_α^2。实际中，往往令$\alpha_i=\alpha+b_i$，此时$\{b_i\}_{i=1}^K$ *i.i.d*均值为 0，方差为$\sigma_b^2=\sigma_\alpha^2$。这样，我们就得到模型：

$y_{ij}=\alpha+\beta x_{ij}+b_i+e_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，

这里$\{b_i\}$ *i.i.d*零均值，方差为σ_b^2，$\{e_{ij}\}$ *i.i.d*为零均值，共同方差为σ_e^2且$\{b_i\}$与$\{e_{ij}\}$相互独立。

此模型中未知的α,β非随机，称为固定效应(fixed effect)，$\{b_i\}$未观察到，为随机效应(random effect)。当然，随机误差$\{e_{ij}\}$也可以称为随机效应。该模型称为线性混合效应模型(linear mixed models, LMM)。

另一种观点是：现在回过头来看我们的分组数据(y_{ij},x_{ij})，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，其中$\sum_{i=1}^K n_i=n$。如果假定这n次观测是独立的，就回到前面建立的线性回归模型：$y_{ij}=\alpha+\beta x_{ij}+r_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，这里$r_{ij}$类似前面的*i.i.d*零均值方差为$\sigma_r^2$随机误差。

在实际中，往往会发现随机误差的变异(方差)σ_r^2所占比重过大。事实上，对不同组中的商品，假定观测独立是比较合理的。但对同一组商品，例如第i组商品，其$1,2,\cdots n_i$次观测应该是非独立的，具有某种相关性。因此固定i，误差$\{r_{ij}\}_{j=1}^{n_i}$独立的假定不是很合理。

为刻画其相关性，且使得模型简单，假设

$$r_{ij}=b_i+e_{ij},$$

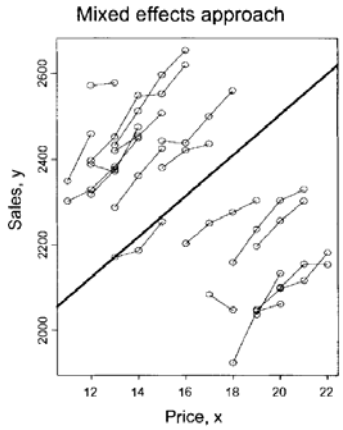
其中，b_i与e_{ij}独立都为零均值的随机变量，方差分别为σ_b^2和σ_e^2。从而$\sigma_r^2=\sigma_b^2+\sigma_e^2$，这也解释了为何以$r_{ij}$作为随机误差，其变异比重过大。组内相关系数

$$\rho=\frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2}。$$

这样，我们也得到线性混合效应模型：

$y_{ij}=\alpha+\beta x_{ij}+b_i+e_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$。

结果如下图：



此时$\rho = 0.99$，表明变异主要是组内的变异。这也解释了为何两种方法估计的斜率符号相反。

再回过头来看我们的分组数据$\left(y_{ij}, x_{ij}\right)$，$i = 1,2,\cdots K$，$j = 1,2,\cdots n_i$，其中$\sum_{i=1}^{K} n_i = n$。对不同组中的商品，假定是独立的。但对同一组商品，例如第i组商品，其每次观测是相关。这样的数据我们成为纵项数据(longitudinal data)，在医学统计中也称为面板数据(panel data)。

一般线性混合效应模型形式为

$$Y = X\beta + Zb + e,$$

其中Y为$n \times 1$观测向量，$X_{n\times p}$为已知设计矩阵，$\beta_{p\times 1}$未知为固定效应，$Z_{n\times q}$为已知设计矩阵，$b_{q\times 1}$为随机效应，且设$Eb = 0$，$Cov(b) = D$非负定，e为随机误差且与b独立，$Ee = 0$，$Cov(e) = R$为正定矩阵。

这样我们得到

$$Cov(Y) = ZDZ' + R。$$

对D，R的不同假设就可以得到不同的线性混合效应模型。

第二章 基础知识补充

2.1 投影矩阵(projection matrix)

定义：矩阵 P称为投影矩阵，若：

1. P是对称的，即$P'=P$；
2. P是幂等的，即$P^2=P$。

幂等矩阵的性质：

1). 特征值非 0 即 1；

2). P幂等则$tr(P)=rank(P)$；

3). P幂等$\Leftrightarrow rank(P)+rank(I_n-P)=n$。

投影矩阵的性质：

1. 投影矩阵是非负定的；
2. 若P_1,P_2是投影矩阵且P_1-P_2非负定，则$P_1P_2=P_2P_1=P_2$，且P_1-P_2也是投影矩阵。

2.2 广义逆

对相容性线性方程

$$A_{m\times n}x=b$$

若 $rank(A)=m=n$ 则方程有唯一解$x=A^{-1}b$。若A不是方阵或者是奇异方阵，Penrose 指出方程的解可以先求解矩阵方程

$$A_{m\times n}B_{n\times m}A_{m\times n}=A_{m\times n}$$

矩阵B称为A的广义逆，记为A^-。

定义 1：对矩阵$A_{m\times n}$，一切满足方程$AXA=A$的矩阵X，称为矩阵A的广义逆，记为A^-，即$AA^-A=A$。

定理 1：设$rank(A_{m\times n})=r$，若A表为

$$A=P_{m\times m}\begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix}Q_{n\times n},$$

其中P,Q可逆，则

$$A^-=Q^{-1}\begin{pmatrix}I_r & B\\ C & D\end{pmatrix}P^{-1}$$

这里B,C,D为适当阶数的任意矩阵。

推论 1：

1. 任何矩阵A的广义逆A^-总存在；
2. A^-唯一$\Leftrightarrow A$可逆；
3. $rank(A^-)\geq rank(A)=rank(A^-A)=rank(AA^-)$

设矩阵$A_{m\times n}$，A的列向量所张成的空间记为$\mu(A)$，即$\mu(A)=\{Ax|x\in R^n\}$。

基本性质：

1.$\mu(A)\subset\mu(B)\Leftrightarrow\exists C,A=BC$；

2.$\dim\mu(A)=rank(A)$。

定理 2：$\mu(A')=\mu(A'A)$。

定理 3：对任何矩阵A

1). $A(A'A)^-A'$ 与 $(A'A)^-$ 的取值无关且$rank\left(A(A'A)^-A'\right)=rank(A)$；

2). $A(A'A)^-A'A=A,\quad A'A(A'A)^-A'=A'$。

定理 4：设线性方程$A_{m\times n}x=b$是相容的，A^-为A给定的一个广义逆，则

1. $x=A^-b$即为方程的一个解；
2. 齐次方程$Ax=0$的所有解为$x=(I_n-A^-A)z$，其中z为任意$n\times1$向量；
3. 方程$Ax=b$的所有解为$x=A^-b+(I_n-A^-A)z$。

推论 4：相容性方程$Ax=b(b\neq0)$的所有解为$\{x|x=A^-b,A^-$为A任一广义逆$\}$。

2.3 正交投影

R^n中两个向量x,y的内积定义为$(x,y)=x'y$，若$x'y=0$，则称$x\perp y$。若$S\subset R^n$为线性子空间，$\forall y\in S,x\perp y$，则称$x\perp S$。令$S^\perp=\{x|x\perp S\}$，则$S^\perp$也是线性子空间，称为S的正交补，易见$S\cap S^\perp=\{0\},S\oplus S^\perp=R^n,(S^\perp)^\perp=S$。

例 1：令$S=\mu(A_{n\times m})\subset R^n,rank(A)=m$，则$S^\perp=\mu(B)$，这里$B=I_n-A(A'A)^{-1}A'$。

设$rank(A_{n\times m})=r$，若$n\times(n-r)$矩阵B满足1.$A'B=0$；2.$rank(B)=n-r$，则称矩阵B为A的正交补，记$B=A^{\perp}$。从定义易见$A^{\perp}$是所有使得$A'B=0$的B秩最大的矩阵。

例 2：设$n\times m$矩阵A，则
$\mu(A^{\perp})=\mu(A)^{\perp},\mu(A^{\perp})\oplus\mu(A)=R^n$。

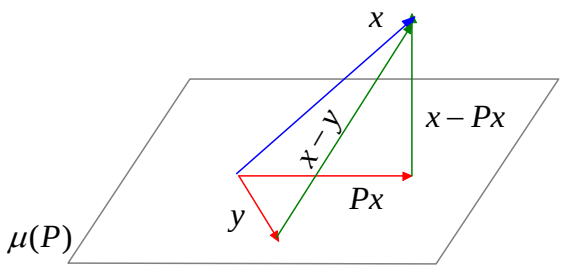
设$x\in R^n$，$S\subset R^n$为线性子空间，x有唯一分解$x=y+z$，$y\in S,z\in S^{\perp}$，称y为x在子空间S上的正交投影(orthogonal projection)。若矩阵$P_{n\times n}$满足对$\forall x\in R^n$，其在子空间S上的正交投影$y=Px$，则称P为S上的正交投影矩阵(简称投影矩阵)。

定理 1：$\mu(A_{n\times m})$上的正交投影矩阵为$P_A=A(A'A)^{-}A'$。

定理 2：P为正交投影矩阵$\Leftrightarrow P$对称幂等。

定理 3：$P_{n\times n}$为投影矩阵$\Leftrightarrow\forall x\in R^n$，$\|x-Px\|=\inf_{y\in\mu(P)}\|x-y\|$。

几何意义



2.4 多元正态分布

定义 1：随机向量$X=(x_1,\cdots,x_n)'$称为n维正态随机向量，如果其密度函数为

$$f(X)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right]$$

这里$\mu=(\mu_1,\cdots,\mu_2)'$，$\Sigma>0$。记$X\sim N_n(\mu,\Sigma)$。

容易计算若$X\sim N_n(\mu,\Sigma)$，则$EX=\mu,Var(X)=\Sigma$。

例 1：二维正态分布密度函数

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\cdot\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

相当于$\mu=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$，$\Sigma=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$。

定义 2：设X为 n维随机向量，若存在非随机的$A_{n\times r}$，$rank(A)=r$以及$\mu_{n\times 1}$使得$X=AU+\mu$，其中$U\sim N_r(0,I_r)$，则称X服从均值为μ、协方差阵为$\Sigma=AA'$的多元正态分布，记为$X\sim N(\mu,\Sigma)$。

此定义把多元正态向量定义为若干个相互独立的一维标准正态分布的线性变换。协方差阵$\Sigma\geq 0$(不一定可逆)。

设$X\sim N(\mu,\Sigma)$，则X的特征函数为

$$\varphi(t)=Ee^{it'X}=\exp\left(it'\mu-\frac{1}{2}t'\Sigma t\right)$$

由于特征函数唯一决定分布函数，因此也可用特征函数来定义多元正态分布，避免$|\Sigma|=0$的情形。

性质：

1.多元正态分布的任意边际分布为相应的多元正态分布；

2.多元正态分布的任线性变换仍然是正态分布，即 $X \sim N(\mu,\Sigma)$，则 $Y = AX + b \sim N(A\mu + b, A\Sigma A')$；

3. 设 $(X_1', X_2')'$ 为多元正态分布，$Cov(X_1, X_2) = 0 \Leftrightarrow X_1, X_2$ 独立；

4. $X \sim N(\mu,\Sigma) \Leftrightarrow \forall t, t'X \sim N(t'\mu, t'\Sigma t)$；

5.设 $X \sim N(\mu,\Sigma)$，$\Sigma > 0$作相应的分块

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

则给定 X_2 时 X_1 的条件分布仍是多元正态分布且条件期望

$$E\left(X_1 \middle| X_2\right) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),$$

条件方差

$$Var\left(X_1 \middle| X_2\right) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}。$$

例 2：设 $(x_1, x_2, x_3, x_4)'$ 联合分布为零均值的正态分布，则

$$Ex_1x_2x_3x_4 = Ex_1x_2Ex_3x_4 + Ex_1x_3Ex_2x_4 + Ex_1x_4Ex_2x_3$$

2.5 正态随机向量的二次型

设 $X_{n\times 1}$ 为随机向量，$A_{n\times n}$ 为对称矩阵，称 $X'AX$ 为随机向量 X 的二次型。

定理 1：设随机向量 X 的均值为 μ，协方差阵为 Σ，则 $E(X'AX) = tr(A\Sigma) + \mu'A\mu$。

定义 1：设 $X \sim N_n(\mu, I_n)$，称随机变量 $y = X'X$ 的分布是服从自由度为 n 的非中心参数为 $\lambda = \mu'\mu$ 的 χ^2 分布，记为 $y \sim \chi^2_{n,\lambda}$；当 $\lambda = 0$ 时称为中心 χ^2 分布记为 $y \sim \chi^2_n$。

$y \sim \chi^2_{n,\lambda}$ 的密度函数为

$$f_\lambda(y) = e^{-\frac{\lambda+y}{2}} \sum_{k=0}^{\infty} \frac{\lambda^k y^{\frac{n}{2}+k-1}}{k!2^{\frac{n}{2}+2k}\Gamma\left(\frac{n}{2}+k\right)}, y > 0。$$

χ^2 分布的基本性质：

1. 设 $y_i \sim \chi^2_{n_i,\lambda_i}$，$i=1,\cdots,k$ 且 y_i 相互独立，则 $\sum_{i=1}^{k} y_i \sim \chi^2_{\sum_{i=1}^{k} n_i, \sum_{i=1}^{k} \lambda_i}$；
2. $E\left(\chi^2_{n,\lambda}\right)=n+\lambda$，$Var\left(\chi^2_{n,\lambda}\right)=2n+4\lambda$；
3. 设 $y \sim \chi^2_{n,\lambda}$，则 y 的特征函数为

$$\varphi(t)=(1-2it)^{-\frac{n}{2}} e^{\frac{i\lambda t}{1-2it}}$$

定理 2：设 $X \sim N_n(\mu, I_n)$，$A_{n\times n}$ 对称，则 $X'AX \sim \chi^2_{r,\mu'A\mu} \Leftrightarrow A$ 幂等且 $rank(A)=r$。

推论 2：设 $X \sim N_n(\mu,\Sigma)$，$\Sigma>0$，$A_{n\times n}$ 对称，则 $X'AX \sim \chi^2_{r,\mu'A\mu}$ 当且仅当下述之一成立

1. $A\Sigma$ 幂等且 $rank(A)=r$；
2. ΣA 幂等且 $rank(A)=r$；
3. Σ 为 A 的一个广义逆且 $rank(A)=r$。

例 3：设 $X \sim N_n(A\beta, \sigma^2 I_n)$，$rank(A)=r$，则 $\frac{X'\left(I_n - A(A'A)^{-}A'\right)X}{\sigma^2} \sim \chi^2_{n-r}$。

定理 3：设 $X \sim N_n(\mu, I_n)$，A，A_1 对称 $X'AX = X'A_1X + X'A_2X \sim \chi^2_{r,\mu'A\mu}$ 且 $X'A_1X \sim \chi^2_{s,\mu'A_1\mu}$，$A_2 \geq 0$ 则：

1. $X'A_2X \sim \chi^2_{r-s,\mu'A_2\mu}$；
2. $X'A_1X$ 与 $X'A_2X$ 独立；
3. $A_1A_2=0$。

推论 3：

1. 设 $X \sim N_n(\mu, I_n)$，A_1, A_2 对称且 $X'A_1X, X'A_2X$ 为 χ^2 分布，则 $X'A_1X, X'A_2X$ 相互独立 $\Leftrightarrow A_1A_2=0$；
2. 设 $X \sim N_n(\mu,\Sigma), \Sigma>0$ A, A_1 对称 $X'AX = X'A_1X + X'A_2X \sim \chi^2_{r,\lambda}$ 且 $X'A_1X \sim \chi^2_{s,\lambda_1}$，$A_2 \geq 0$ 则 $X'A_2X \sim \chi^2_{r-s,\lambda_2}$，$X'A_1X$ 与 $X'A_2X$ 独立且 $A_1\Sigma A_2=0$。

2.6 正态随机向量的线性形式与二次型之间的独立性

设$X \sim N_n(\mu,\Sigma)$，A,B为对称矩阵，C为$m\times n$矩阵，本节研究二次型$X'AX, X'BX$之间以及与线性形式CX之间独立性问题。

定理 1：设$X \sim N_n(\mu, I_n)$，A对称，若$CA=0$，则CX与$X'AX$独立。

推论 1：设$X \sim N_n(\mu,\Sigma),\Sigma>0$，$A$对称，若$C\Sigma A=0$，则$CX$与$X'AX$独立。

例 4：设$x_i \sim N(\mu_i,\sigma^2)$彼此相互独立，令$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$，则$\bar{x}$与$s^2$独立。

定理 2：设$X \sim N_n(\mu, I_n)$，A,B对称，若$AB=0$，则$X'AX$与$X'BX$独立。

推论 2：设$X \sim N_n(\mu,\Sigma),\Sigma>0$，$A,B$对称，若$A\Sigma B=0$，则$X'AX$与$X'BX$独立。

2.7 多元 t 分布(Multivariate t)

一维 t 分布：设x,y独立且$x \sim N(\delta,1)$，$y \sim \chi_n^2$，称$t=x/\sqrt{\frac{y}{n}}$的分布称为自由度为 n 非中心参数为δ的 t 分布，记为$t \sim t_{n,\delta}$；$\delta=0$称为中心的 t 分布；$t \sim t_{n,\delta}$的密度函数为：

$$f(t|n,\delta)=\frac{n^{\frac{n}{2}}e^{-\frac{\delta^2}{2}}}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)(n+t^2)^{\frac{n+1}{2}}}\sum_{i=0}^{\infty}\Gamma\left(\frac{n+i+1}{2}\right)\frac{(\delta t)^i}{i!}\left(\frac{2}{n+t^2}\right)^{\frac{i}{2}}$$

当$\delta=0$时，t_n的密度为

$$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}。$$

定义 1：设随机向量$X_{p\times 1}$有密度

$$f(X)=\frac{\Gamma\left(\frac{n+p}{2}\right)|B|^{\frac{1}{2}}}{(n\pi)^{\frac{p}{2}}\Gamma\left(\frac{n}{2}\right)}\left[1+\frac{(X-\mu)'B(X-\mu)}{n}\right]^{-\frac{n+p}{2}}$$

则称X的分布为p元　t分布或称　t向量，记为$X\sim t_p(\mu,B^{-1},n)$。

注：n为自由度，μ为位置参数，B相当于刻度参数。

例 5：设$X\sim N_p(0,\Sigma)$，$y\sim\chi_n^2$且X与Y独立，则$t=X/\sqrt{\frac{y}{n}}\sim t_p(0,\Sigma,n)$。

第三章：参数估计与分布理论

3.1 最小二乘估计(Least Squared Estimate)

线性模型

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + e_{n\times 1},$$

其中$Ee = 0$，通常对误差 e两种假定：

1.$Cov(e) = \sigma^2 I_n$，σ^2未知；

2.$e \sim N(0, \sigma^2 I_n)$，$\sigma^2$未知(比 1 强)。

由最小二乘法的思想，β的估计应选择使得 $Q(\beta) = \|e\|^2 = \|Y - X\beta\|^2 = (Y - X\beta)'(Y - X\beta)$ 达到最小。

若$\hat{\beta}$使得$\|Y - X\hat{\beta}\|^2 = \min_{\beta}\|Y - X\beta\|^2$，则称$\hat{\beta}$为$\beta$的一个最小二乘解。极小化$Q(\beta)$，则$\beta$需满足方程$\frac{\partial Q(\beta)}{\partial \beta} = 0$，即得到正规方程

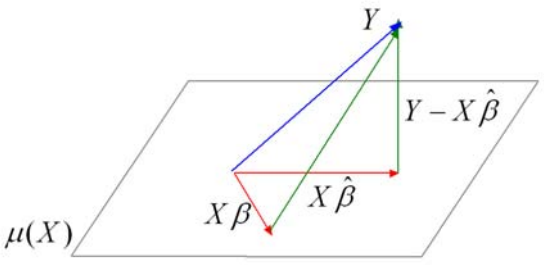
$$X'X\beta = X'Y。$$

正规方程的所有解为$(X'X)^- X'Y$。

定理 3.1.1：$\hat{\beta}$是最小二乘解$\Leftrightarrow\hat{\beta}$是正规方程的解，即$\hat{\beta} = (X'X)^- X'Y$。

最小二乘法几何解释：

$$\min_{\beta}\|Y - X\beta\|^2 = \min_{\theta\in\mu(X)}\|Y - \theta\|^2$$



当$rank(X) = p$时，最小二乘解有唯一解$\hat{\beta} = (X'X)^{-1}X'Y$，且此时$\hat{\beta}$为$\beta$的无偏估计，即$E\hat{\beta} = \beta$，方差$Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$。

当$rank(X) < p$时，最小二乘解不唯一，此时最小二乘解中无 $\hat{\beta}$能作为 β的无偏估计。此外可以证明此时β的无偏估计不存在，此时β称为不可估的(nonestimable)。

定义 3.1.1：$c'\beta$为 β的某一线性函数（c已知），若存在Y的线性函数$a'Y$使得$Ea'Y = c'\beta$，$\forall\beta$，则称 $c'\beta$是可估函数。

可估性的含义可由下面的例子形象说明。

例 1：设两个物体重量β_1,β_2未知，把它们同时放在天平上称 n次，第 i次结果为y_i，模型为

$$y_i=\beta_1+\beta_2+e_i, i=1,\cdots,n.$$

其中e_i为第 i次称量误差，且设$e=(e_1,\cdots,e_n)'$，$Ee=0, Var(e)=\sigma^2 I_n$。

令$c=(1,1)', \beta=(\beta_1,\beta_2)'$，则$c'\beta=\beta_1+\beta_2$是可估的，，例如$y_1$就是其一个无偏估计，但$\beta_1,\beta_2$都不可估。$\beta_1$(或$\beta_2$)不可估的理由很清楚，因为每次称量都是两物体一起称，当然无法由结果对其中单独一个物体的重量作“估计”。

定理 3.1.2：以下三条等价

1.$c'\beta$可估；

2.$X\beta_1=X\beta_2\Rightarrow c'\beta_1=c'\beta_2$；

3.$c\in\mu(X')$。

注：1.一切$c'\beta$可估$\Leftrightarrow rank(X)=p$。

2. 若 $c_1'\beta, c_2'\beta$ 可估，则任线性组合$\lambda_1 c_1'\beta+\lambda_2 c_2'\beta$也是可估。若$c_1,c_2$线性无关，称$c_1'\beta, c_2'\beta$也线性无关。对线性模型来说之多有$rank(X)$个线性无关的可估函数。

3.若$c'\beta$可估，$\hat{\beta}$是最小二乘解，则$c'\hat{\beta}$值唯一，与广义逆$(X'X)^-$的选取无关；此外$c'\hat{\beta}$是$c'\beta$的无偏估计。

定义 3.1.2：设$c'\beta$可估，称$c'\hat{\beta}$为$c'\beta$的最小二乘估计。

例 2：考虑如下模型

$$y_{ij}=\mu+\alpha_i+e_{ij}, i=1,2; j=1,2.$$

$$e_{ij}\ \ i.i.d\ \ Ee_{ij}=0, Ee_{ij}^2=\sigma^2.$$

μ,α_1,α_2不可估，但$\alpha_1-\alpha_2$可估。

若$c'\beta$可估，$a'Y$为其一无偏估计，对$\forall b \in \mu(X)^{\perp}$，$(a+b)'Y$都是$c'\beta$的无偏估计。在所有线性无偏估计中，找出方差最小的估计，此估计称为最优线性无偏估计(Best Linear Unbiased Estimate，简写成 BLUE)，或称为 Gauss-Markov 估计(GM 估计)。

定理 3.1.3：(Gauss-Markov 定理)若$c'\beta$可估，则$c'\hat{\beta}$是其唯一的 GM 估计($\hat{\beta}$为 β的 LS 估计)。

注：在一切线性无偏估计类中，$c'\hat{\beta}$是方差最小的，但不排除比$c'\hat{\beta}$方差更小的非线性无偏估计；但若误差分布还是正态分布，则这种可能性不存在。

下面考虑σ^2的估计。令$\hat{e} = Y - X\hat{\beta} = (I_n - P_X)Y$，称为残差(residual)向量，$E\hat{e} = 0, Cov(\hat{e}) = \sigma^2(I_n - P_X)$。

定理 3.1.4：设$r = rank(X_{n\times p})$，则

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n-r} = \frac{\|Y - X\hat{\beta}\|^2}{n-r} = \frac{Y'(I_n - P_X)Y}{n-r}$$

为σ^2的无偏估计。

3.2 分布理论

迄今为止，关于误差的假定是满足 GM 条件，若误差还服从正态分布即$e \sim N_n(0, \sigma^2 I_n)$，则可以确定一些估计量的精确分布。

定理 3.2.1：在误差正态分布假设下，设$c'\beta$为可估函数，$\hat{\beta} = (X'X)^{-}X'Y$，$rank(X) = r$，则：

1. $c'\hat{\beta}$为$c'\beta$的极大似然估计(MLE)，且$c'\hat{\beta} \sim N\left(c'\beta, \sigma^2 c'(X'X)^{-}c\right)$;

2. $\frac{n-r}{n}\hat{\sigma}^2$是$\sigma^2$的 MLE，且$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-r}$;

3. $c'\hat{\beta}$与$\hat{\sigma}^2$独立。

注：对可估函数来说，其 LSE 与 MLE 一致；但对误差方差的估计，LSE 是无偏的，MLE 是有偏的。

*定理 3.2.2：在正态误差假设下

1.$T_1 = Y'Y$，$T_2 = X'Y$是完全、充分统计量；

2.若$c'\beta$可估，则$c'\hat{\beta}$是唯一最小方差无偏估计(Minimum Variance Unbiased Estimate，简写 MVUE)；

3.$\hat{\sigma}^2$为σ^2的 MVUE。

3.3 有线性约束时的估计

考虑线性模型，其系数 β满足线性约束条件$L\beta = d$(此约束为相容性方程)。不失一般性可假设$d = 0$(否则，取β_0使得$L\beta_0 = d$，令$\tilde{Y} = Y - X\beta_0, \tilde{\beta} = \beta - \beta_0$，考虑线性模型$\tilde{Y} = X\tilde{\beta} + e$，此时线性约束变为$L\tilde{\beta} = 0$)。本节考虑线性模型如下

$$Y = X\beta + e, Ee = 0, Cov(e) = \sigma^2 I_n$$

$$L_{q\times p}\beta = 0$$

由最小二乘法，此时β的估计应选择在$L\beta = 0$的条件下极小化$\|Y - X\beta\|^2$，即选择$\hat{\beta}_L = Arg\min_{L\beta=0}\|Y - X\beta\|^2$。引入 Lagrange 乘子$\lambda_{q\times 1}$，考虑$Q(\beta, \lambda) = \|Y - X\beta\|^2 + 2\lambda' L\beta$，由$\frac{\partial Q}{\partial \beta} = 0, \frac{\partial Q}{\partial \lambda} = 0$得到方程$X'X\beta + L'\lambda = X'Y$，$L\beta = 0$，即

$$\begin{pmatrix} X'X & L' \\ L & 0 \end{pmatrix}\begin{pmatrix} \beta \\ \lambda \end{pmatrix} = \begin{pmatrix} X'Y \\ 0 \end{pmatrix} \qquad (*)$$

引理 3.3.1：设$S = \{A_{n\times m}x \mid B_{k\times m}x = 0, x \in R^m\}$，则 S 为线性子空间且$\dim S = rank\begin{pmatrix} A \\ B \end{pmatrix} - rank(B)$。

引理 3.3.2：设$V_{p\times p} \geq 0$，A为$p\times q$矩阵，则

1.$\mu(A) \cap \mu(VA^{\perp}) = \{0\}$；

2.$\mu(V : A) = \mu(VA^{\perp} : A)$。

回到方程(*)，由$L\beta=0$解得$\beta=(I_p-L^-L)z,\forall z_{p\times 1}$，代入前一个方程得

$$X'X(I_p-L^-L)z+L'\lambda=X'Y \quad (**)$$

由引理 3.3.2，注意到$I_p-L^-L=(L')^{\perp}$，有

$\mu(X')=\mu(X'X)\subset\mu(X'X\vdots L')=\mu\left(X'X(I_p-L^-L)\vdots L'\right)$

。因此方程(**)是相容得，从而方程(*)也是相容的。

定义 3.3.1：若存在a使得$Ea'Y=c'\beta$对所有满足$L\beta=0$的β成立，则称$c'\beta$是条件可估函数，$a'Y$为$c'\beta$的条件无偏估计。

定理 3.3.1：在线性约束$L\beta=0$的条件下，线性函数$c'\beta$是条件可估函数$\Leftrightarrow c\in\mu(X'\vdots L')$。

定理 3.3.2：在线性约束$L\beta=0$的条件下，

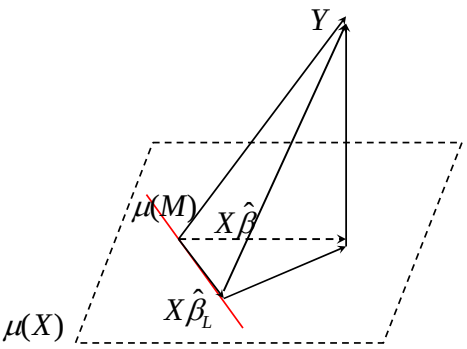
1. 约束最下二乘解为$\hat{\beta}_L=G_{11}X'Y$，其中，

$$\begin{pmatrix} X'X & L' \\ L & 0 \end{pmatrix}^- =\begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix};$$

2. 对任何条件可估函数$c'\beta$，$c'\hat{\beta}_L$值唯一（不依赖广义逆的选取）且$Var(c'\hat{\beta}_L)=\sigma^2c'G_{11}c$；
3. $c'\hat{\beta}_L$是条件可估函数$c'\beta$的唯一 BLUE。

几何解释：

$X\hat{\beta}_L$是Y到子空间$\mu(M)=\{X\beta|L\beta=0\}$的正交投影，这里$M=X(L')^{\perp}=X(I_p-L^-L)$。由于$X\hat{\beta}_L=XG_{11}X'Y$，故$XG_{11}X'$为$\mu(M)$上的投影算子，即$P_M=XG_{11}X'$。



推论 3.3.2：在一些条件下$\hat{\beta}_L$有简单的形式。设$rank(L_{q\times p})=q$，

1. 若$\mu(L')\subset\mu(X')$，则

$$\hat{\beta}_L=\hat{\beta}-(X'X)^{-}L'\left[L(X'X)^{-}L'\right]^{-1}L\hat{\beta}$$

其中$\hat{\beta}=(X'X)^{-}X'Y$；

2. 若$rank(X_{n\times p})=p$，则

$$\hat{\beta}_L=\hat{\beta}-(X'X)^{-1}L'\left[L(X'X)^{-1}L'\right]^{-1}L\hat{\beta}$$

其中$\hat{\beta}=(X'X)^{-1}X'Y$。

定理 3.3.3：在线性约束$L\beta=0$条件下，$\hat{\sigma}_L^2=\dfrac{\left\|Y-X\hat{\beta}_L\right\|^2}{n-s}=\dfrac{Y'(I_n-P_M)Y}{n-s}$为$\sigma^2$的条件无偏估计，这里$M=X(L')^{\perp}$ $s=rank\begin{pmatrix}X\\L\end{pmatrix}-rank(L)$。

定理 3.3.4：在约束$L\beta=0$的条件下，若进一步假设误差还是正态分布，$c'\beta$条件可估，则：

1.$c'\hat{\beta}_L\sim N(c'\beta,\sigma^2c'G_{11}c)$；

2.$\dfrac{(n-s)\hat{\sigma}_L^2}{\sigma^2}\sim\chi^2_{n-s}$，$s=rank\begin{pmatrix}X\\L\end{pmatrix}-rank(L)$；

3.$c'\hat{\beta}_L$与$\hat{\sigma}_L^2$独立。

当 $rank(X_{n\times p})<p$ 时，β 不可估，对 $c\in\mu(X')$，$c'\beta$可估；在带有约束$L\beta=0$的条件下，对$c\in\mu(X'\vdots L')$，$c'\beta$(条件)可估。增加约束后，可估函数的选取范围“扩大”了。加上怎样的约束$L\beta=0$，当然，加上约束后要与原模型一致(等价)，可以使得$\forall c$，$c'\beta$(条件)可估？

设线性模型$Y=X\beta+e, r=rank(X)<p$，由最小二乘法，实质是找 Y在空间$\mu(X)$的投影；加上约束$L\beta=0$后，找 Y在空间 $S=\{X\beta \mid L\beta=0,\beta\in R^p\}$上的投影，要使两模型等价，由于 $S\subset\mu(X)$，要求 $S=\mu(X)\Leftrightarrow \dim S=\dim\mu(X)=rank(X)$，由于$\dim S=rank\begin{pmatrix}X\\L\end{pmatrix}-rank(L)$，所以当

$$rank\begin{pmatrix}X\\L\end{pmatrix}-rank(L)=rank(X)$$

即$\mu(X')\cap\mu(L')=\{0\}$时，加上约束与原模型本质上一样。

另一方面，在约束$L\beta=0$下，要对$\forall c$，$c'\beta$(条件)可估，则 $\mu(X'\vdots L')=R^p$，即 $rank\begin{pmatrix}X\\L\end{pmatrix}=p$。因此，若$L_{q\times p}$满足条件：

$$\mu(X')\cap\mu(L')=\{0\}\text{且}rank\begin{pmatrix}X\\L\end{pmatrix}=p \qquad (\#)$$

则加上约束与不加约束线性模型等价，此时对对$\forall c$，$c'\beta$在条件$L\beta=0$下可估。若 L满足条件(#)，则称约束 $L\beta=0$ 为 side condition(边界条件)。

以下不失一般性，可以假设$L_{q\times p}$是行满秩的，若 L 满足条件(#)，则 $q=p-rank(X)$。

引理 3.3.1：设$L\beta=0$为 side condition，则

$$X'X(X'X+L'L)^{-1}X'=X',$$
$$L(X'X+L'L)^{-1}X'=0。$$

定理 3.3.5：设有约束$L\beta=0$的线性模型，若$L\beta=0$为 side condition，则

1. 约束最小二乘问题的解$\hat{\beta}_L$唯一且$\hat{\beta}_L=(X'X+L'L)^{-1}X'Y$；
2. $\hat{\beta}_L$为方程$\begin{cases}X'X\beta=X'Y\\ L\beta=0\end{cases}$的唯一解；
3. $\hat{\beta}_L$为β在$L\beta=0$条件下的无偏估计且对任何可估函数$c'\beta$ ($c\in\mu(X')$)，$c'\hat{\beta}_L=c'\hat{\beta}$(这里$\hat{\beta}=(X'X)^{-}X'Y$)。

上述定理表明，对通常线性模型，正规方程$X'X\beta=X'Y$的解$\hat{\beta}$一般不唯一。若加上一个 side condition，问题转化求解$\begin{cases}X'X\beta=X'Y\\ L\beta=0\end{cases}$，此时解$\hat{\beta}_L$唯一，且对任可估$c'\beta$，$c'\hat{\beta}_L=c'\hat{\beta}$。Side condition 让线性模型正规方程有一个特殊的解。由于满足 side condition 的L选择并不唯一，$\forall D_{q\times q}$可逆，$DL\beta=0$都是一个 side condition。

可以选择一个“好”的 side condition 使得方程$\begin{cases}X'X\beta=X'Y\\ L\beta=0\end{cases}$解很容易求出。

例 3：考虑下面线性模型：

$$y_{ij}=\mu+\alpha_i+e_{ij},i=1,\cdots,n;j=1,\cdots,m.$$

写成矩阵形式$Y=X_{nm\times(n+1)}\beta+e$，$\beta=(\mu,\alpha_1,\cdots,\alpha_n)'$，$rank\,(X)=n<n+1$，side condition 为一个方程，选择$L\beta=\sum_{i=1}^{n}\alpha_i=0$，易解的$\hat{\mu}=\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}$，$\hat{\alpha}_i=\frac{1}{m}\sum_{j=1}^{m}y_{ij}-\hat{\mu}$。

3.4 Aitken 模型与广义最小二乘

到目前为止，我们都是在误差$Cov(e)=\sigma^2I_n$下讨论估计问题，此时任何可估函数的 LS 估计与其 GM 估计一致。但在许多实际情况下，误差协方差阵为$Cov(e)=\sigma^2\Sigma$，$\Sigma>0$(已知)，在此条件下线性模型$Y=X\beta+e,Ee=0$，称为 Aitken 模型。

由于$\Sigma>0$为已知，则作变换$\tilde{Y}=\Sigma^{-1/2}Y,\tilde{X}=\Sigma^{-1/2}X,\tilde{e}=\Sigma^{-1/2}e$，此时$\tilde{Y}=\tilde{X}\beta+\tilde{e},E\tilde{e}=0,Cov(\tilde{e})=\sigma^2I_n$。

令$Q(\beta)=\left\|\tilde{Y}-\tilde{X}\beta\right\|^2=(Y-X\beta)'\Sigma^{-1}(Y-X\beta)$，由最小二乘法原理，此时$\beta$的估计应选择$\beta^*=Arg\min_{\beta}Q(\beta)$。此时得到 Aitken 方程：

$$X'\Sigma^{-1}X\beta=X'\Sigma^{-1}Y,$$

解$\beta^*=\left(X'\Sigma^{-1}X\right)^{-}X'\Sigma^{-1}Y$，称为广义最小二乘解。

注：如果Σ为对角阵$diag(\sigma_{11},\sigma_{22},\cdots\sigma_{nn})$，令$W=\Sigma^{-1}=diag(w_1,w_2,\cdots w_n)$，这里$w_i=\sigma_{ii}^{-1}$，此时$\beta^*$也称加权最小二乘估计。

由于可估性只涉及Y的均值与其方差无关，且$\mu(\tilde{X})=\mu(X')$，故此时$c'\beta$可估$\Leftrightarrow c\in\mu(X')$。

注：由于$X\left(X'\Sigma^{-1}X\right)^{-}X'$不依赖于广义逆的选取，令$A=X\left(X'\Sigma^{-1}X\right)^{-}X'\Sigma^{-1}$，由于$AX=X$，因此$A^2=A$为幂等矩阵。

定理 3.4.1：对任何可估函数$c'\beta$，$c'\beta^*$是唯一的 BLUE，其方差$Var(c'\beta^*)=\sigma^2c'\left(X'\Sigma^{-1}X\right)^{-}c$。

定理 3.4.2：$\sigma^{*2}=\dfrac{\left(Y-X\beta^*\right)'\Sigma^{-1}\left(Y-X\beta^*\right)}{n-r}$，$r=rank(X)$为$\sigma^2$的无偏估计。

定理 3.4.3：若假设$e\sim N(0,\sigma^2\Sigma),\Sigma>0$，则

1. 对任何可估$c'\beta$，$c'\beta^*\sim N\left(c'\beta,\sigma^2c'(X'\Sigma^{-1}X)^{-}c\right)$;

2. $\dfrac{(n-r)\sigma^{*2}}{\sigma^2}\sim\chi^2_{n-r}$且$c'\beta^*$与$\sigma^{*2}$独立;

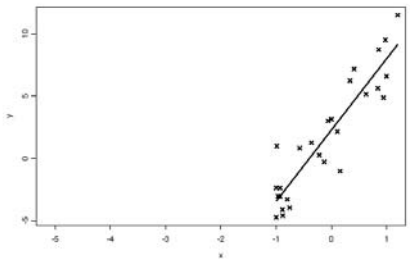
3. 若$rank(X_{n\times p})=p$，则

$\beta^*\sim N_p\left(\beta,\sigma^2(X'\Sigma^{-1}X)^{-1}\right)$且与$\sigma^{*2}$独立。

3.5 稳健回归与M-估计

到目前为止，我们对线性模型都采用最小二乘法给出回归系数的参数估计。前面也指出，在误差$e\sim N(0,\sigma^2I_n)$条件下，回归系数的最小二乘估计还是极大似然估计。但也很多证据表明当误差分布偏离正态分布时，最小二乘估计可能不再是有效的估计(估计的方差会偏大)。即便误差正态分布，如果有些异常值影响，最小二乘法都会过于敏感，给出很“坏”的估计。

例 3.5.1:下图显示异常值对线性模型最小二乘估计的影响。



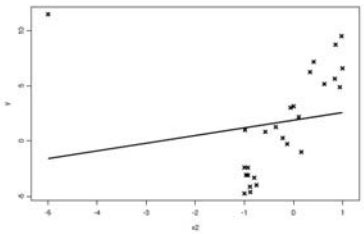


Figure *Data with 27 points and the corresponding least squares regression line (top) and the sensitivity of least squares regression to an outlier in the x-direction (bottom).*

Huber(1964)首先考虑位置参数的稳健(robust)估计，提出一类现在称为M-估计的估计方法。对线性模型来说，给定一合适的函数$\rho(\cdot)$，考虑最小化问题

$$\min_{\beta}\sum_{i=1}^{n}\rho(y_i-x_i'\beta),$$

上式最小化问题的解$\hat{\beta}_n$，称为回归系数β的M-估计。

注：最小二乘法相当于$\rho(x)=x^2$。

如果，函数ρ的导数$\varphi=\dot{\rho}$处处存在且连续，则M-估计$\hat{\beta}_n$为下面方程的一解

$$\sum_{i=1}^{n}x_i\varphi(y_i-x_i'\beta)=0。$$

注：在一些情况下，导函数$\varphi=\dot{\rho}$不连续或者函数ρ在一些至多可数的点不存在，即$\varphi=\dot{\rho}$除去这些点存在，此时上方程可能无解。此时，M-估计的定义只能回到原始的极值问题。

M-估计的一个常用特例是取L_1范数，对应函数$\rho(x)=|x|$，此时最小化

$$\min_{\beta}\sum_{i=1}^{n}|y_i-x_i'\beta|,$$

其解$\hat{\beta}_n$，称为 least absolute deviation（LAD）估计。研究表明，最小二乘法之所以不稳健是因为$\rho(x)=x^2$，$|\varphi(x)|=|\dot{\rho}(x)|=2|x|$，当$|x|$过大时增长太快，会“放大”异常值的影响。而 LAD 方法是稳健的方法，因为$\rho(x)=|x|$，$|\varphi(x)|=|\dot{\rho}(x)|=1$(在 0 处导数不存在)，增长平缓。

Huber(1964)提出一类函数，现在称为 Huber 函数，设$\varphi=\dot{\rho}$，取

$$\varphi_H(x)=\begin{cases}x & |x|\le k\\ k\,sign(x) & |x|>k\end{cases},$$

此时得到的估计称为 *Huber* 估计。

一些人建议减少异常值影响，当$x\to\pm\infty$时$\varphi(x)\to 0$，由此可取$\varphi(x)=\dfrac{2x}{1+x^2}$。

注：关于M-估计的渐近理论可参见陈希孺，赵林城(1996)，线性模型中的 M 方法, 上海科学技术出版社。

3.6*最小二乘统一理论

对线性模型：

$Y=X\beta+e$，$Ee=0, Cov(e)=\sigma^2\Sigma$($\Sigma$已知)

若$\Sigma>0$，按照广义最小二乘法，只需求解关于参数β的二次函数$Q(\beta)=(Y-X\beta)'\Sigma^{-1}(Y-X\beta)$的极小值问题；

若$|\Sigma|=0$，此时称为奇异线性模型，由于Σ^{-1}不存在，$Q(\beta)$无定义，若用广义逆Σ^{-}代替Σ^{-1}，把$Q(\beta)$定义为$Q(\beta)=(Y-X\beta)'\Sigma^{-}(Y-X\beta)$，

则由于$Q(\beta)$与广义逆Σ^{-}的选择有关，不同的Σ^{-}，$Q(\beta)$不同，因此极小化$Q(\beta)$无意义。因此对于奇异线性模型，一个核心问题是寻找一个新矩阵T，能够充当$Q(\beta)$中Σ^{-1}所担负的作用。C.R.Rao 成功的解决了这个问题，他定义

$T=\Sigma+XUX'$，其中$U\ge 0$，

且使得$rank(T)=rank(\Sigma\vdots X)$，然后令$Q(\beta)=(Y-X\beta)'T^{-}(Y-X\beta)$，$\beta$的估计应选择$\beta_R^*=Arg\min_{\beta}Q(\beta)$。

后面将证明，对于可估函数$c'\beta$，其 BLUE 就是$c'\beta_R^*$。而且此方法适用于设计矩阵X列满秩或不满秩，Σ奇异或非奇异的所有情形，因此此方法和结果称为最小二乘的统一理论。

引理 3.6.1：对本节线性模型，$Y \in \mu(\Sigma \vdots X)$以概率 1 成立。

引理 3.6.2：对上述定义的 T，

1. $\mu(\Sigma \vdots X) = \mu(T)$；

2. $X'T^-X, X'T^-Y$和$Q(\beta)$不依赖广义逆 T^-的选取。

引理 3.6.3：对线性模型，可估函数$c'\beta$的某一个无偏估计$a'Y$为其 BLUE⇔对任意0的无偏估计$b'Y$总有$Cov(a'Y, b'Y) = 0$。

注：满足 Rao 定义的 T总是存在的，例如取 $U = k \cdot I_n (k > 0)$ 。特别当 $\Sigma > 0$ 或 $\mu(X) \subset \mu(\Sigma)$时，可以取$U = 0$，此时$T = \Sigma$。

极小化 $Q(\beta)$ 得相容性方程 $X'T^-X\beta = X'T^-Y$，解$\beta_R^* = \left(X'T^-X\right)^- X'T^-Y$。

定理 3.6.1：对本节线性模型

1. 对任何可估函数$c'\beta$，其 BLUE 为$c'\beta_R^*$；

2. $Var(c'\beta_R^*) = \sigma^2 c'[\left(X'T^-X\right)^- - U]c$；

3. σ^2的无偏估计为

$$\sigma_R^{*2} = \frac{\left(Y - X\beta_R^*\right)' T^- \left(Y - X\beta_R^*\right)}{q}$$

这里$q = rank(T) - rank(X)$。

最后作为本章结束提一下两步估计方法。当$Cov(e) = \sigma^2\Sigma(\theta)$，其中$\Sigma(\theta) > 0$含有未知参数 θ，先设法对 θ作估计，得到 $\hat{\theta}$，从而得到$\Sigma(\theta)$的估计$\hat{\Sigma}(\hat{\theta})$，最后用广义最小二乘的思想得到 β 的估计 $\hat{\beta}(\hat{\theta}) = \left(X'\left[\hat{\Sigma}(\hat{\theta})\right]^{-1}X\right)^- X'\left[\hat{\Sigma}(\hat{\theta})\right]^{-1}Y$。对于可估函数 $c'\beta$，在一定条件下$c'\hat{\beta}(\hat{\theta})$为 $c'\beta$ 的无偏估计。

第四章：假设检验与区间估计

4.1 F检验

考虑如下线性模型：

$$Y = X_{n\times p}\beta + e,\ e \sim N(0,\sigma^2 I_n),$$

设$rank(X)=r$，矩阵$H_{m\times p}$(已知)，线性假设

$$H_0: H\beta = 0 \leftrightarrow H_1: H\beta \neq 0,$$

不失一般性设$rank(H)=m$，现要检验假设H_0(作出拒绝或接受该假设的判断)。

考虑该假设的似然比(likelihood ratio)检验。设似然函数

$$L(Y;\beta,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left(-\frac{\|Y-X\beta\|^2}{2\sigma^2}\right),$$

则似然比定义为

$$\lambda = \frac{\sup\limits_{\beta,\sigma^2} L(Y;\beta,\sigma^2)}{\sup\limits_{\substack{\beta,\sigma^2 \\ H\beta=0}} L(Y;\beta,\sigma^2)}。$$

注意到当$\hat\beta = (X'X)^- X'Y$，$\tilde\sigma^2 = \frac{\|Y-X\hat\beta\|^2}{n}$时似然比的分子达到最大，且$\sup\limits_{\beta,\sigma^2} L(Y;\beta,\sigma^2) = \left(\sqrt{\frac{2\pi e}{n}}\|Y-X\hat\beta\|\right)^{-n}$，当$\hat\beta_H$为约束$H\beta=0$下最小二乘估计，$\tilde\sigma_H^2 = \frac{\|Y-X\hat\beta_H\|^2}{n}$时似然比的分母达到最大，且$\sup\limits_{\substack{\beta,\sigma^2 \\ H\beta=0}} L(Y;\beta,\sigma^2) = \left(\sqrt{\frac{2\pi e}{n}}\|Y-X\hat\beta_H\|\right)^{-n}$。

因此似然比

$$\lambda = \left(\frac{\|Y-X\hat\beta_H\|^2}{\|Y-X\hat\beta\|^2}\right)^{\frac{n}{2}} = \left(1+\frac{k}{n-r}F\right)^{\frac{n}{2}},$$

这里

$$F = \frac{(ESS_H - ESS)/k}{ESS/(n-r)},$$

$$ESS = \|Y-X\hat\beta\|^2,\ ESS_H = \|Y-X\hat\beta_H\|^2,$$

$$k = rank(X) + rank(H) - rank\begin{pmatrix}X\\H\end{pmatrix}。$$

ESS——error sum of squares(误差平方和)

由定义似然比$\lambda \geq 1$，直观上若H_0成立，则λ应充分接近 1，故若λ偏离 1 很远则应该拒绝H_0，由于λ是F的单调函数，故当F很大时应该拒绝H_0。

定理 4.1.1：在本节线性模型假设下，

1. $\frac{ESS}{\sigma^2} \sim \chi^2_{n-r}$；
2. $\frac{ESS_H - ESS}{\sigma^2} \sim \chi^2_{k,\delta}$，这里 $\delta = \frac{\left\| X(\beta - E\hat{\beta}_H) \right\|^2}{\sigma^2}$；
3. $ESS_H - ESS$与ESS独立；
4. 在假设H_0下，$F \sim F_{k,n-r}$。

当$\mu(H') \subset \mu(X')$，即$H\beta$的每个分量都是可估的，此时$rank\begin{pmatrix} X \\ H \end{pmatrix} = rank(X)$，因此$k = rank(H) = m$，$F = \frac{(ESS_H - ESS)/m}{ESS/(n-r)}$在假设$H_0$下服从$F_{m,n-r}$分布。由定理 3.3.2 的推论，此时

$\hat{\beta}_H = \hat{\beta} - (X'X)^- H' \left[H(X'X)^- H' \right]^{-1} H\hat{\beta}$。

$$ESS_H - ESS = \left\| X(\hat{\beta}_H - \hat{\beta}) \right\|^2 = (H\hat{\beta})' \left[H(X'X)^- H' \right]^{-1} (H\hat{\beta})$$

给定水平$\alpha \in (0,1)$，若$F > F_{m,n-r}(\alpha)$，则拒绝假设H_0：$H\beta = 0$。此检验称为 F-检验。

关于F统计量的另一种形式。由于$ESS = Y'Y - \hat{\beta}'X'Y$，$ESS_H = Y'Y - \hat{\beta}_H'X'Y$，记$RSS = \hat{\beta}'X'Y$，$RSS_H = \hat{\beta}_H'X'Y$，则$F = \frac{(RSS - RSS_H)/m}{ESS/(n-r)}$。

RSS ——regression sum of squares(回归平方和)

TSS ——total sum of squares(总和)

令$TSS = Y'Y$，则有$TSS = RSS + ESS$。

下面从投影矩阵的角度来看F统计量。线性模型可表示为$Y = \theta + e, e \sim N(0, \sigma^2 I_n)$，其中$\theta \in \mu(X)$。在假设$H_0$下(设$\mu(H') \subset \mu(X')$)，模型可表为$Y = \theta + e, e \sim N(0, \sigma^2 I_n)$，其中$\theta \in S = \{X\beta | H\beta = 0, \beta \in R^p\}$，注$S \subset \mu(X)$。

$\dim S = rank\begin{pmatrix} X \\ H \end{pmatrix} - rank(H) = r - m$。记 P，P_S 分别为子空间$\mu(X)$，S上的投影矩阵，线性模型θ的最小二乘估计为$\hat{\theta} = PY$，在假设下的最小二乘估计为$\hat{\theta}_H = P_S Y$，由于$\theta \in \mu(X)$，$(I_n - P)\theta = 0$，因此

$SSE = \left\| Y - \hat{\theta} \right\|^2 = Y'(I_n - P)Y = e'(I_n - P)e$。

在假设H_0下的线性模型，$\theta \in S$，$(I_n - P_S)\theta = 0$，因此 $ESS_H = \left\| Y - \hat{\theta}_H \right\|^2 = Y'(I_n - P_S)Y = e'(I_n - P_S)e$。由定理 3.2.1，$\frac{ESS}{\sigma^2} \sim \chi^2_{n-r}$，在假设 H_0 下 $\frac{ESS_H}{\sigma^2} \sim \chi^2_{n+m-r}$，$ESS_H - ESS = e'(P - P_S)e$，由于$(P - P_S)(I - P) = 0$，故 $ESS_H - ESS$与 ESS独立且$\frac{ESS_H - ESS}{\sigma^2} \sim \chi^2_m$。因此在 H_0下，$F \sim F_{m,n-r}$。

4.2 有初始约束时的假设检验

设有初始约束的线性模型

$$\begin{cases} Y = X\beta + e, e \sim N(0, \sigma^2 I_n) \\ L\beta = 0, rank(L_{q \times p}) = q \end{cases}$$

考虑假设H_0: $H\beta = 0$的检验问题，这里$H\beta$为m个条件可估函数，即$\mu(H') \subset \mu(X' \vdots L')$。记 $\hat{\beta}_L$，$\hat{\beta}_{LH}$ 分别为 β 在约束 $L\beta = 0$ 和 $L\beta = 0, H\beta = 0$ 的最小二乘估计，由定理 3.3.2，$\hat{\beta}_L = G_{11} X'Y$(参见$G_{11}$的定义)。

$\hat{\beta}_{LH}$ 可以简单由$\begin{pmatrix} L \\ H \end{pmatrix}$代替$\hat{\beta}_L$中的$L$得到。

从而由残差平方和：

$$ESS_L = \left\| Y - X\hat{\beta}_L \right\|^2 = Y'Y - \hat{\beta}_L' X'Y,$$

$$ESS_{LH} = \left\| Y - X\hat{\beta}_{LH} \right\|^2 = Y'Y - \hat{\beta}_{LH}' X'Y。$$

此时投影子空间分别为：

$$S = \left\{ X\beta \middle| L\beta = 0, \beta \in R^p \right\},$$

$$S_1 = \left\{ X\beta \middle| L\beta = 0, H\beta = 0, \beta \in R^p \right\}。$$

$$\dim S = rank\begin{pmatrix} X \\ L \end{pmatrix} - rank(L) \equiv m_1,$$

$$\dim S_1 = rank\begin{pmatrix} X \\ L \\ H \end{pmatrix} - rank\begin{pmatrix} L \\ H \end{pmatrix} \equiv m_2。$$

此时假设H_0的似然比检验F-统计量为：

$$F = \frac{(ESS_{LH} - ESS_L)/(m_1 - m_2)}{ESS_L/(n - m_1)},$$

当 H_0成立时$F \sim F_{m_1 - m_2, n - m_1}$分布。

4.3 一般线性假设检验

至此都是讨论假设检验为$H\beta = 0$，对一般线性假设检验$H_{m\times p}\beta = d$，$rank(H) = m$且方程$H\beta = d$是相容的，先考虑线性假设$H\beta$每个分量都是可估的，即$\mu(H') \subset \mu(X')$。

设$rank(X) = r$，β_0为方程$H\beta = d$某一特解即 $H\beta_0 = d$，对线性模型 $Y = X\beta + e$，$e \sim N(0, \sigma^2 I_n)$，令$Z = Y - X\beta_0$，$\theta = \beta - \beta_0$，得到线性模型：

$$Z = X\theta + e，\ e \sim N(0, \sigma^2 I_n)，$$

此时，对该模型要检验的假设变为$H\theta = 0$。因此对于假设 $H\beta = d$，$rank(H) = m$，$\mu(H') \subset \mu(X')$，检验的F-统计量为

$$F = \frac{\left(H\hat{\beta} - d\right)'\left[H(X'X)^{-}H'\right]^{-1}\left(H\hat{\beta} - d\right)/m}{\left\| Y - X\hat{\beta} \right\|^2/(n - r)},$$

在假设$H\beta = d$的条件下$F \sim F_{m,n-r}$，给定水平$\alpha \in (0,1)$，若$F > F_{m,n-r}(\alpha)$，则拒绝该假设。

更一般的，若线性假设$H\beta$包含有不可估函数，即$\mu(H') \not\subset \mu(X')$，不失一般性，把$H$剖分成$H=\begin{pmatrix}H_1\\H_2\end{pmatrix}$，其中$H_1\beta$可估，$H_2\beta$不可估。考虑线性模型

$$Y=\theta+e,\quad \theta\in\mu(X),\quad e\sim N(0,\sigma^2 I_n)$$

的如下两个假设检验问题：

$$H_0:\ H\beta=0,$$

$$H_{01}:\ H_1\beta=0。$$

从正交投影来看，两个假设相当于

$$H_0:\ \theta\in S=\{X\beta|H\beta=0,\beta\in R^p\},$$

$$H_{01}:\ \theta\in S_1=\{X\beta|H_1\beta=0,\beta\in R^p\}。$$

显然有$S\subset S_1$，另一方面$H_1\beta$可估，$H_2\beta$不可估，因此

$$\begin{aligned}\dim S_1&=rank\begin{pmatrix}X\\H_1\end{pmatrix}-rank(H_1),\\&=rank(X)-rank(H_1)\end{aligned}$$

$$\begin{aligned}\dim S&=rank\begin{pmatrix}X\\H\end{pmatrix}-rank(H)\\&=rank\begin{pmatrix}X\\H_2\end{pmatrix}-rank\begin{pmatrix}H_1\\H_2\end{pmatrix},\\&=rank(X)-rank(H_1)\end{aligned}$$

故$S=S_1$。这表明假设$H_0:\theta\in S$与假设$H_{01}:\theta\in S_1$是一样的。因此对于不可估的假设是无法检验的，称为不可检验的假设(non-testable hypothesis)。

例 4.3.1：检验两样本是否同一线性模型

$$y_i=\beta_0^{(1)}+\beta_1^{(1)}x_{i1}+\cdots\beta_{p-1}^{(1)}x_{i,p-1}+e_i,i=1,\cdots,n_1,$$

$$y_i=\beta_0^{(2)}+\beta_1^{(2)}x_{i1}+\cdots\beta_{p-1}^{(2)}x_{i,p-1}+e_i,i=n_1+1,\cdots,n_1+n_2,$$

其中误差$i.i.d\sim N(0,\sigma^2)$，要检验两组数据是否来于同一模型，即检验$\beta_i^{(1)}=\beta_i^{(2)},0\le i\le p-1$。

首先写成矩阵形式，有两个线性模型$i=1,2$

$Y_i=X_i\beta_i+e_i$，$e_i\sim N_{n_i}(0,\sigma^2 I_{n_i})$，$rank(X_i)=p$

要检验 $H_0:\ \beta_1=\beta_2$。

将原问题写成一个线性模型：

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

要检验假设H_0：$\begin{pmatrix} I_p & -I_p \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = 0$。

在假设H_0下，$\beta_1 = \beta_2 = \beta$，模型为$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$，最小二乘估计

$$\hat{\beta}_H = (X_1'X_1 + X_2'X_2)^{-1}(X_1'Y_1 + X_2'Y_2)。$$

记$\hat{\beta}_i = (X_i'X_i)^{-1}X_i'Y_i, i = 1,2$，则检验的$F$统计量为$F = \dfrac{(ESS_H - ESS)/p}{ESS/(n_1 + n_2 - 2p)}$，其中

$ESS_H = Y_1'Y_1 + Y_2'Y_2 - \hat{\beta}_H'(X_1'Y_1 + X_2'Y_2)$，

$ESS = Y_1'Y_1 + Y_2'Y_2 - (\hat{\beta}_1'X_1'Y_1 + \hat{\beta}_2'X_2'Y_2)$。

在假设H_0下，$F \sim F_{p,n_1+n_2-2p}$分布，给定水平$\alpha \in (0,1)$，若$F > F_{p,n_1+n_2-2p}(\alpha)$，则拒绝该假设。

4.4 置信椭球(confidence ellipsoid)

设线性模型$Y = X_{n\times p}\beta + e$，$e \sim N(0, \sigma^2 I_n)$，$rank(X) = r$，$\Phi = H_{m\times p}\beta = \begin{pmatrix} h_1'\beta \\ \vdots \\ h_m'\beta \end{pmatrix}$为$m$个独立的可估函数，即$rank(H) = m$，$\mu(H') = \mu(X')$。

令$\hat{\beta} = (X'X)^- X'Y$，则$\hat{\Phi} = H\hat{\beta}$为$\Phi$的BLUE，且$\hat{\Phi} \sim N_m(\Phi, \sigma^2 V)$，这里$V = H(X'X)^- H' > 0$。由推论 3.2.2

$$\frac{(\hat{\Phi} - \Phi)'V^{-1}(\hat{\Phi} - \Phi)}{\sigma^2} \sim \chi_m^2。$$

由定理 3.2.1，σ^2的估计$\hat{\sigma}^2 = \dfrac{\|Y - X\hat{\beta}\|^2}{n - r}$且与$\hat{\Phi}$独立，$\dfrac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-r}^2$，从而

$$\frac{(\hat{\Phi}-\Phi)'V^{-1}(\hat{\Phi}-\Phi)}{m\hat{\sigma}^2}\sim F_{m,n-r}。$$

因此对$\forall\alpha\in(0,1)$，

$$P\left(\frac{(\hat{\Phi}-\Phi)'V^{-1}(\hat{\Phi}-\Phi)}{m\hat{\sigma}^2}\le F_{m,n-r}(\alpha)\right)=1-\alpha。$$

令

$$D=\left\{\Phi\middle|(\Phi-\hat{\Phi})'V^{-1}(\Phi-\hat{\Phi})\le m\hat{\sigma}^2F_{m,n-r}(\alpha)\right\},$$

是以$\hat{\Phi}$为中心的一个椭球，$P(\Phi\in D)=1-\alpha$，D称为Φ的置信系数为$1-\alpha$的置信椭球。

用$H\beta$，$H\hat{\beta}$，$H(X'X)^-H'$代替$\Phi,\hat{\Phi},V$，则置信椭球可写为

$$\left(H\beta-H\hat{\beta}\right)'\left[H(X'X)^-H'\right]^{-1}\left(H\beta-H\hat{\beta}\right)\le m\hat{\sigma}^2F_{m,n-r}(\alpha)。$$

特别若$m=1$，由于$F_{1,n-r}$与$t^2{}_{n-r}$分布一致，令$t_{n-r}(\alpha/2)$为上$\alpha/2$分位点，则此时可估函数$h'\beta$的$1-\alpha$置信区间为：

$$h'\hat{\beta}\pm t_{n-r}(\alpha/2)\hat{\sigma}\sqrt{h'(X'X)^-h},$$

或$h'\hat{\beta}\pm t_{n-r}(\alpha/2)\hat{\sigma}_{h'\hat{\beta}}$，其中$\hat{\sigma}^2_{h'\hat{\beta}}=\hat{\sigma}^2h'(X'X)^-h$为$Var(h'\beta)$的估计。

4.5 同时置信区间与 Bonferroni t-区间

往往要对若干个可估函数同时给出区间估计。设有m个可估函数$\phi_1=h_1'\beta,\cdots,\phi_m=h_m'\beta$，用上一节的方法可以对每个$\phi_i$作一个置信水平为$1-\alpha$的$t$-区间估计$\hat{\phi}_i\pm t_{n-r}(\alpha/2)\hat{\sigma}_{\hat{\phi}_i}$，$1\le i\le m$。由不等式

$$P\left(\bigcap_{i=1}^m A_i\right)=1-P\left(\bigcup_{i=1}^m\overline{A}_i\right)\ge1-\sum_{i=1}^m P(\overline{A}_i)，\quad(*)$$

若每个事件发生的概率$P(A_i)=1-\alpha$，则

所有事件同时发生的概率不是$1-\alpha$，而是只能保证$P\left(\bigcap_{i=1}^m A_i\right)\ge1-m\alpha$。例如$\alpha=0.05$，$m=10$，则只能保证$\ge0.5$。一般来说虽然每个置信区间置信系数是$1-\alpha$，同时置信区间的置信系数比$1-\alpha$要低。若要确保 m 个联合置信区间同时成立的概率达到名义上的$1-\alpha$，一个可供选择的办法是把每个置信区间的置信系数提高到$1-\frac{\alpha}{m}$。

上述作法当m不太大($m\le 5$)，效果还可以，但总的来说过于保守，特别当 m很大时，此时每个置信区间都太宽以致没有多大的实际意义。切合实际的折衷办法是增大α。

由于不等式(*)称为 Bonferroni 不等式，基于用$\frac{\alpha}{m}$替换 α的方法得到的同时置信区间 $h_i'\hat{\beta}\pm t_{n-r}(\alpha/2m)\hat{\sigma}_{h_i'\hat{\beta}}, i=1,\cdots,m$ 称为 Bonferroni t-区间。

4.6 最大模 t-区间

对 m个独立可估函数$h_1'\beta,\cdots,h_m'\beta$作同时区间估计，令 $H_{m\times p}=\begin{pmatrix} h_1' \\ \vdots \\ h_m' \end{pmatrix}$，$rank(H)=m$，$V=(v_{ij})_{m\times m}=H(X'X)^{-}H'$，则$H\hat{\beta}\sim N_m(H\beta,\sigma^2V)$，$H\hat{\beta}$ 与 $\hat{\sigma}^2$ 独立且 $\frac{(n-r)\hat{\sigma}^2}{\sigma^2}\sim\chi^2_{n-r}$ 分布。令 $x_i=\frac{h_i'\hat{\beta}-h_i'\beta}{\sqrt{v_{ii}}}$，

$X=(x_1,\cdots,x_m)'$，则 $X\sim N_m(0,\sigma^2R)$，其中 $R=(r_{ij})_{m\times m}$，$r_{ij}=\frac{v_{ij}}{\sqrt{v_{ii}v_{jj}}}$。因此令$t=(t_1,\cdots,t_m)'$，$t_i=\frac{h_i'\hat{\beta}-h_i'\beta}{\hat{\sigma}\sqrt{v_{ii}}}$，则

$$t\sim t_m(0,R,n-r)(\text{多元 } t\text{-分布})。$$

令$t_m^{\alpha/2}(0,R,n-r)$使得

$$P\left(-t_m^{\alpha/2}(0,R,n-r)\le t_i\le t_m^{\alpha/2}(0,R,n-r),1\le i\le m\right)=1-\alpha$$

即 $P\left(\max_{1\le i\le m}|t_i|\le t_m^{\alpha/2}(0,R,n-r)\right)=1-\alpha$。

这样 m个区间

$$h_i'\hat{\beta}\pm\hat{\sigma}\sqrt{v_{ii}}t_m^{\alpha/2}(0,R,n-r),i=1,\cdots,m$$

为置信系数$1-\alpha$的同时置信区间。由于$t_m^{\alpha/2}(0,R,n-r)$是由 m个 t分布变量取最大模分布确定的，故上同时置信区间称为最大模 t 区间(maximum modulus t-intervals)。

计算上区间的关键是求出$t_m^{\alpha/2}(0,R,n-r)$，一般是比较困难。Sidak(1968)证明了

$$t_m^{\alpha/2}(0,R,n-r)\le t_m^{\alpha/2}(0,I_m,n-r),$$

因此

$$P\left(h_i'\beta\in h_i'\hat{\beta}\pm\hat{\sigma}\sqrt{v_{ii}}t_m^{\alpha/2}(0,I_m,n-r),1\le i\le m\right)\ge 1-\alpha,$$

而$t_m^{\alpha/2}(0,I_m,n-r)$是易求出的。

4.7 Scheffe 区间和置信带

引理 4.7.1：设$A_{n\times n}>0$，则$\sup\limits_{b\neq 0}\dfrac{(a'b)^2}{b'Ab}=a'A^{-1}a$。

定理 4.7.1：设线性模型$Y=X\beta+e$，$e\sim N(0,\sigma^2 I_n)$，$rank(H_{m\times p})=m$，$\mu(H')\subset\mu(X')$，则对任意可估$l'\beta$，$l\in\mu(H')$，其置信系数为$1-\alpha$的同时置信区间为

$$l'\hat{\beta}\pm\left[mF_{m,n-r}(\alpha)\right]^{1/2}\hat{\sigma}\left[l'(X'X)^{-}l\right]^{1/2}。$$

上述同时置信区间是 Scheffe(1953)提出来的，称为 Scheffe 区间。注意 Scheffe 区间不是有限多个可估函数的同时置信区间，它是所有$l'\beta$，$l\in\mu(H')$的同时置信区间。对有限可估函数同时置信区间，Scheffe 方法不一定最好，其长度可能会偏长，但 Scheffe 方法可以用于所有线性模型而无需对设计矩阵做任何限制。

当$m=r$即$\mu(H')=\mu(X')$时，可以得到所有可估函数$l'\beta$的同时$1-\alpha$置信区间。

若$rank(X_{n\times p})=p$，此时任何线性函数$l'\beta$都是可估的，此时

$$P\left(l'\beta\in l'\hat{\beta}\pm\left[pF_{p,n-p}(\alpha)\right]^{1/2}\hat{\sigma}\left[l'(X'X)^{-}l\right]^{1/2},\forall l\in R^p\right)=1-\alpha。$$

当 l变化时，区间

$$l'\hat{\beta}\pm\left[pF_{p,n-p}(\alpha)\right]^{1/2}\hat{\sigma}\left[l'(X'X)^{-}l\right]^{1/2}$$

也变化，形成一个区域，称为置信带(confidence band)，其宽度为$2\left[pF_{p,n-p}(\alpha)\right]^{1/2}\hat{\sigma}\left[l'(X'X)^{-}l\right]^{1/2}$。

例 4.7.1：设简单线性模型

$y_i=\beta_0+\beta_1x_i+e_i$，$i=1,\cdots,n$，$e_i\sim N(0,\sigma^2)$独立同分布。$\hat{\beta}=\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$，其中

$\hat{\beta}_1=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})/\sum_{i=1}^{n}(x_i-\bar{x})^2$，

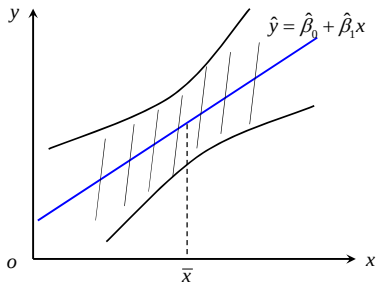
$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$，$\bar{x}=\sum_{i=1}^{n}x_i/n$，$\bar{y}=\sum_{i=1}^{n}y_i/n$，

$\hat{\sigma}^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2/(n-2)$。

对任线性函数$\beta_0+\beta_1x$，其$1-\alpha$置信带为

$(\hat{\beta}_0+\hat{\beta}_1x)\pm\left[2F_{2,n-2}(\alpha)\right]^{1/2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$。

置信带关于经验直线$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$对称，当$x=\bar{x}$时带宽最小。



4.8 预测问题

假定响应变量 y与协变量$x_1,\cdots,x_p$存在线性关系$y=\beta_1x_1+\cdots+\beta_px_p+e$，设有 n次观测，则未知参数$\beta_1,\cdots,\beta_p$可以由前面的结果来作出估计，设估计为$\hat{\beta}_1,\cdots,\hat{\beta}_p$，这样得到经验模型$\hat{y}=\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p$，该经验模型近似的描述了响应变量 y与协变量$x_1,\cdots,x_p$之间的关系，给定协变量的值，可以对响应变量作出预测。

点预测

考虑线性模型$y_i = x_i'\beta + e_i$，$i = 1,\cdots,n$，写成矩阵形式：

$Y = X\beta + e$，$Ee = 0$，$Cov(e) = \sigma^2\Sigma$，$rank(X_{n\times p}) = r$，$\Sigma > 0$(已知)。现有m个点$x_i = (x_{i1},\cdots,x_{ip})'$，$i = n+1,\cdots,n+m$，感兴趣的问题是由此预测响应变量$y$的$m$个值$y_{n+1},\cdots,y_{n+m}$。

令$X_0 = \begin{pmatrix} x_{n+1}' \\ \vdots \\ x_{n+m}' \end{pmatrix}$，$Y_0 = \begin{pmatrix} y_{n+1} \\ \vdots \\ y_{n+m} \end{pmatrix}$，$e_0 = \begin{pmatrix} e_{n+1} \\ \vdots \\ e_{n+m} \end{pmatrix}$，则$Y_0 = X_0\beta + e_0$，$Ee_0 = 0$，$Cov(e_0) = \sigma^2\Sigma_0$。假设$\mu(X_0') \subset \mu(X')$。

首先考虑被预测量Y_0与历史数据Y不相关，此时$Cov(e_0, e) = 0$，为预测Y_0，一个直观的方法就是用$EY_0 = X_0\beta$的估计作为预测。即用

$$Y_0^* = X_0\beta^* = X_0(X'\Sigma^{-1}X)^- X'\Sigma^{-1}Y$$

来预测Y_0，这里β^*为广义最小二乘估计。由于$\mu(X_0') \subset \mu(X')$，所以$Y_0^*$与广义逆的选取无关。预测的偏差$Z = Y_0^* - Y_0$，则$EZ = 0$，即预测$Y_0^*$为$Y_0$的一个无偏估计。偏差的协方差阵$Cov(Z) = \sigma^2\left[\Sigma_0 + X_0(X'\Sigma^{-1}X)^- X_0'\right]$。

以上传统预测方法假定被预测量Y_0与历史数据Y不相关，但在一些情形Y_0与Y相关。

设Y_0与Y相关程度$Cov(e_0, e) = \sigma^2 V_{m\times n}$($V$已知)，设$Y_0^* = C_{m\times n}Y$为$Y_0$的一个线性预测，评价预测$Y_0^*$好坏目前常用的度量是广义预测均方误差(generalized prediction mean squared error，简写 PMSE)。给定$A > 0$，$PMSE(Y_0^*) = E(Y_0^* - Y_0)'A(Y_0^* - Y_0)$。若线性预测是无偏的且广义预测均方误差最小，则称该线性预测是最优线性无偏预测(best linear unbiased predictor，简写 BLUP)。

定理 4.8.1：对于本节线性模型，若 $Cov(e_0,e)=\sigma^2V_{m\times n}$（$V$已知)且$X_0\beta$可估，则

$$Y_0^*=X_0\beta^*+V\Sigma^{-1}(Y-X\beta^*)$$

为Y_0的最优线性无偏预测。特别若 $V=0$，则Y_0的最优线性无偏预测为$Y_0^*=X_0\beta^*$。

区间预测

以下假设误差为正态分布，即 $e\sim N_n(0,\sigma^2\Sigma)$，$e_0\sim N_n(0,\sigma^2\Sigma_0)$，为简单起见，只考虑 Y与 Y_0不相关情形，即 $V=0$。在正态误差假设下，偏差$Z=Y_0^*-Y_0$服从正态分布 $N\left(0,\sigma^2\left[\Sigma_0+X_0(X'\Sigma^{-1}X)^-X_0'\right]\right)$。设 $\mu(X_0')\subset\mu(X')$，$rank(X_{n\times p})=r$，$\sigma^{*2}=\dfrac{(Y-X\beta^*)'\Sigma^{-1}(Y-X\beta^*)}{n-r}$，$\Sigma_0=\left(\sigma_{ij}^{(0)}\right)_{n+1\le i,j\le n+m}$，

则类似 4.5 节，对每个$i=n+1,\cdots,n+m$，y_i的$1-\alpha$预测区间为

$x_i'\beta^*\pm t_{n-r}(\alpha/2)\sigma^*\left[\sigma_{ii}^{(0)}+x_i'(X'\Sigma^{-1}X)^-x_i\right]^{1/2}$，

利用 Bonferroni 方法，$y_{n+1},\cdots,y_{n+m}$的一个置信系数至少$1-\alpha$同时预测区间为

$x_i'\beta^*\pm t_{n-r}(\alpha/2m)\sigma^*\left[\sigma_{ii}^{(0)}+x_i'(X'\Sigma^{-1}X)^-x_i\right]^{1/2}$，

$n+1\le i\le n+m$。

也可以由 Scheffe 方法得到$y_{n+1},\cdots,y_{n+m}$的一个置信系数至少$1-\alpha$的同时预测区间：

$x_i'\beta^*\pm\left[mF_{m,n-r}(\alpha)\right]^{1/2}\sigma^*\left[\sigma_{ii}^{(0)}+x_i'(X'\Sigma^{-1}X)^-x_i\right]^{1/2}$，

$n+1\le i\le n+m$。

例 4.8.1：简单线性回归模型 $y_i=\beta_0+\beta_1x_i+e_i$，$e_i\sim N(0,\sigma^2)$独立同分布。现感兴趣的是同时预测$y_{n+1},\cdots,y_{n+m}$(设相互独立)。对每个$i=n+1,\cdots,n+m$，$y_i$的点预测为：$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$。

$y_{n+1},\cdots,y_{n+m}$的一个置信系数至少$1-\alpha$同时预测 Bonferroni 区间为：

$$(\hat{\beta}_0+\hat{\beta}_1 x_i)\pm t_{n-r}\left(\frac{\alpha}{2m}\right)\hat{\sigma}\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{k=1}^{n}(x_k-\bar{x})^2}\right]^{1/2}$$

，$n+1\le i\le n+m$。

$y_{n+1},\cdots,y_{n+m}$的一个置信系数至少$1-\alpha$同时预测 Scheffe 区间为：

$$(\hat{\beta}_0+\hat{\beta}_1 x_i)\pm\left[mF_{m,n-r}(\alpha)\right]^{1/2}\hat{\sigma}\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{k=1}^{n}(x_k-\bar{x})^2}\right]^{1/2}$$

，$n+1\le i\le n+m$。

第五章：模型及诊断

5.1 含常数项的线性模型

前面研究的都是一般形式下的线性模型。在实际中，往往线性模型含有常数项(截距项)。设此时线性模型为

$$y=\beta_0+\beta_1x_1+\cdots+\beta_{p-1}x_{p-1}+e,$$

这里β_0为常数项，$x_1,x_2\cdots x_{p-1}$为协变量，e为随机误差。若记$x_0=1$，$x=(x_0,x_1,\cdots x_{p-1})'$，$\beta=(\beta_0,\beta_1,\cdots\beta_{p-1})'$，则即为通常线性模型情形$y=x'\beta+e$。

设有n次观测，写成矩阵形式，令

$$Y=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix},\quad X=\begin{pmatrix}1 & x_{11} & \cdots & x_{1,p-1}\\ 1 & x_{21} & \cdots & x_{2,p-1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{n,p-1}\end{pmatrix},$$

$$\beta=\begin{pmatrix}\beta_0\\ \beta_I\end{pmatrix}=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_{p-1}\end{pmatrix},\quad e=\begin{pmatrix}e_1\\ \vdots\\ e_n\end{pmatrix},\ \text{则}$$

$$Y=X\beta+e$$

假设$Ee=0$，$Cov(e)=\sigma^2I_n$，$rank(X_{n\times p})=p$。最小二乘估计$\hat\beta=(X'X)^{-1}X'Y$，令$\hat\sigma^2=\dfrac{\|Y-X\hat\beta\|^2}{n-p}$。

在上述假定下，将第三章中的结论搬过来有

1.$E\hat\beta=\beta$，$Var(\hat\beta)=\sigma^2(X'X)^{-1}$，$E\hat\sigma^2=\sigma^2$；

2. (Gauss-Markov 定理)对$\forall c'\beta$，$c'\hat\beta$是其唯一的 BLUE；

若进一步假定误差为正态分布，则

3.对$\forall c'\beta$，$c'\hat\beta$是其唯一的 MVUE；

4.$\hat\beta\sim N_p\left(\beta,\sigma^2(X'X)^{-1}\right)$，$\dfrac{(n-p)\hat\sigma^2}{\sigma^2}\sim\chi^2_{n-p}$，$\hat\beta$与$\hat\sigma^2$独立。

对此模型，主要感兴趣的是回归系数β_I的估计，常数项β_0单独考虑。令$E_{n\times1}=(1,\cdots,1)'$，$X_{n\times p}=\left(E_n\vdots\tilde X_{n\times(p-1)}\right)$，则模型为$Y=\beta_0E_n+\tilde X\beta_I+e$。

在实际应用中，对数据中心化是常用的手段。所谓中心化就是把自变量的度量起点移至到n次试验中所取值的中心点处。记

$\bar{x}_j=\dfrac{\sum_{i=1}^{n}x_{ij}}{n},1\le j\le p-1$，$\bar{x}=(\bar{x}_1,\cdots,\bar{x}_{p-1})'$，

$\bar{y}=\dfrac{\sum_{i=1}^{n}y_i}{n}$，则中心化后模型分量形式为：

$$y_i=\alpha+\beta_1(x_{i1}-\bar{x}_1)+\cdots+\beta_{p-1}(x_{i,p-1}-\bar{x}_{p-1})+e_i$$

其中$\alpha=\beta_0+\bar{x}'\beta_I$，写成矩阵形式为

$Y=\alpha E_n+\tilde{X}_c\beta_I+e$，$Ee=0$，$Cov(e)=\sigma^2I_n$，

其中$\tilde{X}_c=\left(I_n-\dfrac{E_nE_n'}{n}\right)\tilde{X}$。$\tilde{X}_c$称为中心化了的设计矩阵，易见$\tilde{X}_c'E_n=0$。此时线性回归模型称为中心化的线性回归模型。正规方程：

$$\begin{pmatrix}n & 0\\ 0 & \tilde{X}_c'\tilde{X}_c\end{pmatrix}\begin{pmatrix}\alpha\\ \beta_I\end{pmatrix}=\begin{pmatrix}n\bar{y}\\ \tilde{X}_c'Y\end{pmatrix}$$

解得

$$\hat{\alpha}=\bar{y},\quad \hat{\beta}_I=(\tilde{X}_c'\tilde{X}_c)^{-1}\tilde{X}_c'Y,$$

$$Cov\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}_I\end{pmatrix}=\sigma^2\begin{pmatrix}1/n & 0\\ 0 & (\tilde{X}_c'\tilde{X}_c)^{-1}\end{pmatrix}。$$

中心化的线性模型，常数项由样本均值估计，回归系数β_I的估计等价于线性回归模型$Y=\tilde{X}_c\beta_I+e$的参数估计。若误差正态分布，则中心化后的模型估计$\hat{\alpha}$与$\hat{\beta}_I$独立。

定理 5.1.1：中心化后给出的回归系数估计与没有中心化时给出的估计是一致的。

除了中心化，对协变量经常作另一种处理。

令 $s_j^2=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$，$1\le j\le p-1$，

$Z=(z_{ij})_{n\times(p-1)}$，其中$z_{ij}=\dfrac{x_{ij}-\bar{x}_j}{s_j}$，则$\sum_{i=1}^{n}z_{ij}^2=1$。

z_{ij}是将x_{ij}中心化后再标准化，易见$E_n'Z=0$。

令$R=(r_{ij})_{(p-1)\times(p-1)}=Z'Z$，则

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{s_i s_j}$$

若把协变量看成随机的，则 r_{ij}正好是协变量x_i与x_j的样本相关系数。中心化后标准化的好处在于：

1.R 可以分析协变量之间的相关关系；

2.消去了单位和取值范围的差异(R无量纲)。

用Z作为设计矩阵，此时分量形式为：$1\le i\le n$，

$y_i=\alpha^{(0)}+\frac{x_{i1}-\bar{x}_1}{s_1}\beta_1^{(0)}+\cdots\frac{x_{i,p-1}-\bar{x}_{p-1}}{s_{p-1}}\beta_{p-1}^{(0)}+e_i$。

这里$\alpha^{(0)}=\alpha$，$\beta_i^{(0)}=s_i\beta_i$，$1\le i\le p-1$。记$\beta_I^{(0)}=\left(\beta_1^{(0)},\cdots,\beta_{p-1}^{(0)},\right)'$，写成矩阵形式：

$$Y=\alpha^{(0)}E_n+Z\beta_I^{(0)}+e,$$

最小二乘估计

$$\hat{\alpha}^{(0)}=\bar{y},\quad \hat{\beta}_i^{(0)}=s_i\hat{\beta}_i,\quad 1\le i\le p-1。$$

5.2 哑(或虚拟)变量(dummy variable)处理

在实际应用中，常常会遇到一些协变量为属性变量，设其属性有k个状态，固然可以用数字$1,2\cdots k$来标识，但不可用来计算，因为它们无数量意义。解决办法是引进哑变量(dummy variable)，$x_{(1)},x_{(2)},\cdots x_{(q)}$，其中$q=k-1$，

$$x_{(i)}=\begin{cases}1,\text{若处在}i\text{状态}\\0,\qquad\text{其它}\end{cases},\quad i=1,2,\cdots q.$$

故$x_{(1)}=x_{(2)}=\cdots x_{(q)}=0$表示处在$k$状态。

若对响应变量y只考察此一因素，模型为

$Ey=\beta_0+\beta_1x_{(1)}+\cdots+\beta_qx_{(q)}$，则易见

$E(y|\text{状态}j)=\beta_0+\beta_j,\ j=1,2,\cdots q.$

$E(y|\text{状态}k)=\beta_0$，故此取哑变量的方法是以状态k为标准，β_j衡量状态j超出状态k的值。

另一种取法是$x_{(j)}=\begin{cases}1,\text{若处在状态}j\\-1,\text{若处在状态}k\\0,\qquad\text{其它}\end{cases}$

此时$x_{(1)}=x_{(2)}=\cdots x_{(q)}=-1$表示状态$k$。因而

$$E(y|\text{状态}j)=\beta_0+\beta_j,\ j=1,2,\cdots q.$$

$$E(y|\text{状态}k)=\beta_0-(\beta_1+\beta_2\cdots+\beta_q)$$

于是$\dfrac{\sum_{j=1}^{k}E(y|\text{状态}j)}{k}=\beta_0$，故$\beta_0$为平均效应，而$\beta_1,\cdots,\beta_q$衡量状态超出平均的效应。

5.3 显著性检验

对回归系数作出估计后就可以得到经验回归方程。所建立的经验回归方程是否真正地刻画了响应变量与协变量之间的实际依赖关系呢？

对线性回归模型：$1\le i\le n$，

$y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_{p-1}x_{i,p-1}+e_i$，$e_i\sim N(0,\sigma^2)$。

首先考虑响应变量 y是否线性地依赖协变量$x_1,\cdots,x_{p-1}$这个整体，即检验假设

$$H_0:\ \beta_1=\cdots=\beta_{p-1}=0。$$

上检验称为回归方程的显著性检验。若假设H_0被接受，意味着相对误差 e而言，所有协变量对响应变量 Y的影响是不重要的。

将模型中心化，写成矩阵形式：

$$Y=\alpha E_n+\tilde{X}_c\beta_I+e=(E_n\vdots\tilde{X}_c)\begin{pmatrix}\alpha\\ \beta_I\end{pmatrix}+e,\ e\sim N(0,\sigma^2I_n)。$$

要检验的假设为

$$H_0:\ H\begin{pmatrix}\alpha\\ \beta_I\end{pmatrix}=0，\text{其中}H_{(p-1)\times p}=(0\vdots I_{p-1})。$$

该假设可以由F检验来给出拒绝域。具体地

$$F=\frac{\hat{\beta}_I'\tilde{X}_c'Y/(p-1)}{(Y'Y-n\bar{y}^2-\hat{\beta}_I'\tilde{X}_c'Y)/(n-p)},$$

在假设H_0下，$F\sim F_{p-1,n-p}$，故给定水平$\alpha\in(0,1)$，当$F>F_{p-1,n-p}(\alpha)$时拒绝假设H_0，否则接受H_0。

当回归方程的显著性检验结果是拒绝原假设时，仅说明至少有一个$\beta_j\neq0$，并不排除响应变量 y不依赖其中某些协变量。

于是在整体的回归方程显著性检验被拒绝后还需对每个自变量逐一地作显著性检验，即对固定的某个 i，作如下假设检验H_i：$\beta_i = 0$。

对线性模型$Y = X\beta + e$，$e \sim N(0, \sigma^2 I_n)$，估计$\hat{\beta} \sim N_p(\beta, \sigma^2 (X'X)^{-1})$，$\hat{\sigma}^2 = \dfrac{\|Y - X\hat{\beta}\|^2}{n-p}$，令$C = (c_{ij})_{p\times p} = (X'X)^{-1}$，则$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$。

在假设H_i下，$t_i = \dfrac{\hat{\beta}_i}{\sqrt{c_{ii}}\hat{\sigma}} \sim t_{n-p}$，故给定水平$\alpha$，当$|t_i| > t_{n-p}(\alpha/2)$时拒绝$H_i$，否则接受$H_i$。

若经过检验，接受原假设H_i：$\beta_i = 0$，认为协变量 x_i对响应变量 y无显著影响，因而可以将其从回归方程中剔除，此时 y对剩余协变量重新作回归，回归系数的估计也随之变化，然后再检验剩余回归系数是否为零，再剔除经检验对y无显著影响的协变量，这样的过程一直下去。

5.4 回归协变量的选择

通常在作回归分析(以后若不特别指明，假设模型都含有常数项，即为 5.1 节中的形式)时，根据问题本身的专业理论及有关经验，常常把各种与响应变量有关或可能有关的协变量引入到回归模型。其结果是把一些对响应变量影响很小，甚至无影响的协变量都选入回归模型中，不但计算量大，而且估计和预测的精度也会下降。

此外，在一些情况下，某些协变量观测数据的获得代价昂贵，若这些协变量对响应变量影响很小或根本没有影响，若不加选择的引进回归模型，势必造成观测数据收集和模型应用的费用不必要的增大。

因此，对模型协变量的精心选择是十分有必要的。

设响应变量 y以及一系列的协变量$x_1, \cdots, x_s$以及这些量的 n次观测值，要识别哪些协变量 x_j对响应变量 y是重要的。

定义 5.4.1：令$R^2=\frac{RSS}{TSS}=\frac{\hat{\beta}_I'\tilde{X}_c'Y}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$，称 R^2 为决定系数(coefficient of determination)。

注 1：在线性回归分析中，β_I是关注的焦点。一般线性模型总和是$Y'Y$，在回归分析中常数项的回归平方和为$n\bar{y}^2$，因此这里总和实际上是去掉常数项的回归平方和，即$TSS=Y'Y-n\bar{y}^2$。

注 2：由等式 $TSS=RSS+ESS$，$(ESS=\left\|Y-X\hat{\beta}\right\|^2)$因此$0\le R^2\le 1$。$R^2$反映了回归和在总和所占的比例，$R^2$越大，表示回归协变量解释的越好。

注 3：将协变量看成随机的，则 y与$(x_1,\cdots,x_{p-1})'$ 的复相关系数 (multiple correlation coefficient)定义为

$$\rho=\sqrt{\frac{Cov(y,x)Var(x)^{-1}Cov(x,y)}{Var(y)}}$$

$\frac{1}{n}\tilde{X}_c'(Y-\bar{y}E_n)$，$\frac{1}{n}\tilde{X}_c'\tilde{X}_c$，$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$ 分别作为$Cov(x,y),Var(x),Var(y)$相应的样本估计，这样得到复相关系数 ρ的估计

$$\hat{\rho}=\sqrt{\frac{\hat{\beta}_I'\tilde{X}_c'Y}{\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

恰好为R，所以 R称为(样本)复相关系数。

注 4：F统计量与R^2的关系：$F=\frac{R^2}{1-R^2}\cdot\frac{n-p}{p-1}$。

若回归协变量个数固定时，则应选择 R^2大的那个回归。但当协变量个数不一样时，用R^2来选择协变量就失效了，因为全部变量都作为协变量，R^2的值将达到最大。

Adjusted R^2 criterion (调整 R^2准则)：回归协变量集的选择应使得$Adj\ R^2$ 达到最大，其中$Adj\ R^2=1-\frac{n-1}{n-p}(1-R^2)$，$p$为协变量的个数(含常数项)。

同样的道理，若回归协变量个数固定时，则应选择误差平方和ESS小的那个回归。但当协变量个数不一样时，用ESS来选择协变量就失效了，因为全部变量都作为协变量，此时ESS的值将达到最小，故必须对协变量的个数加一个“惩罚因子”。令

$$RMS_p = \frac{ESS}{n-p} = \frac{\|Y - X\beta\|^2}{n-p} = \hat{\sigma}^2$$

RMS_p criterion(平均残差平方和准则，residual mean squares criterion)：回归协变量集的选择应使 RMS_p达到最小。

令$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$为 y的样本方差，则

$Adj\ R^2 = 1 - \frac{RMS_p}{s_y^2}$，故$Adj\ R^2$准则与$RMS_p$准则本质上是一样的。

上面的准则是从回归拟合角度来看，Mallows'C_p准则则是从预测角度出发。设使用所有协变量的完全模型(含常数项)为：

$$Y = X_{n\times s}\beta + e,$$

给定一个含有 p个参数的子集模型(含常数项)，得到经验方程$\hat{Y} = X_p\hat{\beta}_p$，用 $\hat{Y}$去预测EY，其均方误差为：

$$MSE_p = E(\hat{Y} - EY)'(\hat{Y} - EY)$$
$$= (X_p\beta_p - EY)'(X_p\beta_p - EY) + p\sigma^2$$

去掉刻度的影响，即无量纲化，令

$$J_p = \frac{MSE_p}{\sigma^2} = \frac{(X_p\beta_p - EY)'(X_p\beta_p - EY)}{\sigma^2} + p,$$

从预测角度来看，应该选择J_p最小的那个回归子集。但J_p不是统计量，因此要找到J_p的一个合理估计。由于对于子集模型，

$$E(ESS) = E(Y - \hat{Y})'(Y - \hat{Y})$$
$$= (EY - X_p\beta_p)'(EY - X_p\beta_p) + (n-p)\sigma^2$$

从而有

$$J_p = \frac{E(ESS)}{\sigma^2} - n + 2p,$$

因此定义 $C_p = \frac{ESS}{\hat{\sigma}^2} - n + 2p$，其中$\hat{\sigma}^2$为全模型的方差估计，即$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}_s\|^2}{n-s}$，统计量$C_p$为$J_p$的一个合理估计。

C_p criterion：回归协变量集的选择应使 C_p 达到最小。

Akaike Information Criterion(AIC 准则)

设$y_1, \cdots, y_n$为一组样本，服从某个含p个参数的模型，参数用向量$\theta_{p\times 1}$表示，似然函数为$l_p(Y|\theta)$，设参数 θ的极大似然估计为$\hat{\theta}$，令

$$AIC_p = -2\log l_p(Y|\hat{\theta}) + 2p,$$

AIC 准则：(回归)模型应选择使统计量AIC_p达到最小的一组参数。

对正态线性模型，具体地有似然函数

$$l(\beta_p, \sigma^2|Y) = (2\pi\sigma^2)^{-n/2}\exp-\frac{\|Y - X_p\beta_p\|^2}{2\sigma^2},$$

极大似然估计 $\beta_p^* = (X_p'X_p)^{-1}X_p'Y$，$\sigma^{*2} = \frac{ESS}{n}$，其中$ESS = \|Y - X_p\beta_p^*\|^2$，略去与$p$无关的常数得到

$$AIC_p = n\log ESS + 2p,$$

回归协变量集的选择应使得AIC_p达到最小的那个。

以上无论哪一种回归协变量的选择准则都需要对不同协变量的子集进行比较，计算量相当大。

5.5 回归诊断与 Box-Cox 变换

到目前为止得到的估计、检验及其他分析都是在认为模型以及关于模型的假设都是正确的情况下得到的。在许多实际问题中，有时这些关于模型的假设是令人怀疑的，需要作一些诊断。

回归诊断关心的是两个相互有关的问题。首先是模型在多大程度上与观测数据相一致；其次令人感兴趣的问题是，每个案例在估计及综合分析的影响。在某些数据集中，若一个案例被删除，估计或分析的统计量可能有重要改变，这样一个案例称为有影响性的(强影响点)，需要检测出这样的案例并分析原因。

设线性回归模型(含截距项)

$$Y = X\beta + e,\ Ee = 0,\ Cov(e) = \sigma^2 I_n$$

$rank(X_{n\times p}) = p$，X 的第一列为 $E_n = (1,\cdots,1)'$。

或者写成中心化形式 $Y = \alpha E_n + \tilde{X}_c\beta_I + e = \left(E_n \vdots \tilde{X}_c\right)\begin{pmatrix}\alpha\\ \beta_I\end{pmatrix} + e$（见 5.1 节）

β 的估计为 $\hat{\beta} = (X'X)^{-1}X'Y$，对应观测值 Y 的拟合值为 $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$，其中 $H = X(X'X)^{-1}X'$ 称为“帽子矩阵”(hat matrix)。

$H = (h_{ij})_{p\times p}$ 的性质：

1. H 对称幂等，$rank(H) = tr(H) = p$；
2. $HX = X$，$HE_n = E_n$。

令 $\hat{e} = Y - X\hat{\beta} = (I_n - H)Y$，称为残差向量(residual vector)，有性质：

1. $E\hat{e} = 0, Cov(\hat{e}) = \sigma^2(I_n - H)$；
2. $Cov(\hat{e}, \hat{Y}) = 0$；
3. $E_n'\hat{e} = 0$。

记 $\hat{e}=(\hat{e}_1,\cdots,\hat{e}_n)'$，其中 $\hat{e}_i=y_i-x_i'\hat{\beta}$，则 $Var(\hat{e}_i)=\sigma^2(1-h_{ii})$。若 $h_{ii}\approx 1$，则 $Var(\hat{e}_i)=Var(y_i-\hat{y}_i)\approx 0$，由于 $E(\hat{e}_i)=E(y_i-\hat{y}_i)=0$，因此 $y_i-\hat{y}_i\approx 0$。

表明点 (x_i',y_i)，其实际值 y_i 与理论值 $\hat{y}_i$ 拟合的特别好，或者说，这样的点有把拟合的回归平面拖向自己的倾向。这样的点常成为**高杠杆点**(high leverage point)。

进一步，由于

$$h_{ii}=(1\quad x_i')\left(X'X\right)^{-1}\begin{pmatrix}1\\ x_i\end{pmatrix}$$
$$=\frac{1}{n}+(x_i-\bar{x})'\left(\tilde{X}_c'\tilde{X}_c\right)^{-1}(x_i-\bar{x})$$

这里，$\bar{x}=\frac{1}{n}E_n'X$ 为数据中心。若点 x_i 距数据中心越远，则 h_{ii} 值越大，越有将回归平面拉向自己的倾向，如下图。

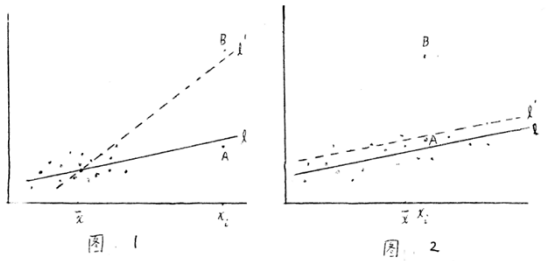
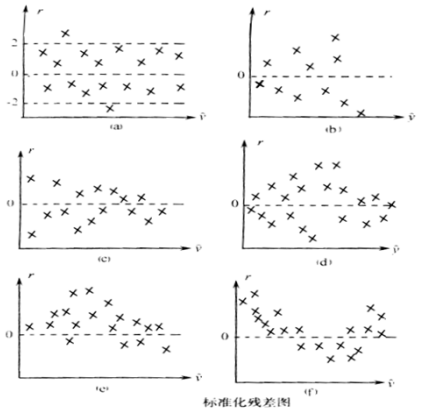


图 1，x_i 远离中心 $\bar{x}$，其 $h_{ii}\approx 1$ 为高杠杆点。若 (x_i,y_i) 在 A 处(属于正常)，则尚不能引起回归直线的基本走向。若在 B 处，则把回归直线由 l 拖向 l'(拉向自己)。好像有一个力将 l 的右端以 $(\bar{x},\bar{y})$ 为定点的杠杆抬上去。这正式高杠杆点这一名称的由来。

而图 2，x_i 离中心 $\bar{x}$ 近，非高杠杆点。此时即便 (x_i,y_i) 处在离群位置 B 处，其作用把回归直线由 l 平行的带上去一点点到 l'。意味着不会较大影响回归系数的估计，仅仅影响了截距项的估计。

若误差正态分布，则$\hat{e}_i \sim N(0,\sigma^2(1-h_{ii}))$，标准化$\frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$，由于 σ^2未知，用 $\hat{\sigma}^2$代替，令$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$，称为“学生化内残差”(因为$\hat{\sigma}^2$用到第 i个案例在内的全部数据)，由于r_i与$\hat{y}_i$轻微相关，故分布比较复杂，但可以近似的认为 r_i相互独立服从$N(0,1)$分布。将点$(\hat{y}_i, r_i)$，$i=1,\cdots,n$描在平面上就可以得到残差图。



线性模型剔除第 i组数据后，剩余$n-1$组数据线性回归模型记为

$Y_{(i)} = X_{(i)}\beta + e_{(i)}$，$Ee_{(i)} = 0$，$Cov(e_{(i)}) = \sigma^2 I_{n-1}$。

β 的 LS 估计记为$\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)}$，为刻画第 i组数据对回归系数估计影响定义

$$D_i = \frac{(\hat{\beta}-\hat{\beta}_{(i)})' X'X (\hat{\beta}-\hat{\beta}_{(i)})}{p\hat{\sigma}^2}$$

称为 Cook 距离。

关于D_i

1. D_i的等值线是椭圆，与置信椭圆具有相同形状；

2.通过线性变换改变X的列，D_i的值不变；

3. 令 $\hat{Y} = X\hat{\beta}$ ， $\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$ ， 则 $D_i = \frac{(\hat{Y}-\hat{Y}_{(i)})'(\hat{Y}-\hat{Y}_{(i)})}{p\hat{\sigma}^2}$刻画了$\hat{Y}$与 $\hat{Y}_{(i)}$的距离。

D_i 大的案例对 $\hat{\beta}$ 及对拟合值 $\hat{Y}$ 都有实质性的影响，删除它们可能会导致结论的重大改变，这样的点称为强影响点。

定理 5.5.1：$D_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}} r_i^2$。

用 Cook 统计量给出判定强影响点的临界值是困难的，在实际中要视具体情况而定。

若一个案例不遵从某个模型，但其余数据遵从，则该案例称为异常值(outliers)。产生异常值的原因一般有二：

1. 用线性模型来近似实际，可能在一定范围是比较好的，当超出该范围时，会产生异常而不适合线性模型；
2. 由于变量测量误差或者数据不正确。

假设第 i 个案例可能是异常值，则

1. 从数据中删除第 i 个案例，余下的 $n-1$ 个案例拟合线性模型；
2. 使用删除后的数据集估计 β 和 σ^2，记为 $\hat{\beta}_{(i)}$ 和 $\hat{\sigma}^2_{(i)}$，以表示第 i 个案例没有用于估计(注 $\hat{\sigma}^2_{(i)}$ 的自由度为 $n-p-1$)；
3. 对于被删除的案例，计算其拟合值 $\hat{y}_{(i)} = x_i'\hat{\beta}_{(i)}$，由于第 i 个案例没有用于估计，y_i 与 $\hat{y}_{(i)}$ 相互独立，

方差 $Var(y_i - \hat{y}_{(i)}) = \sigma^2\left[1 + x_i'\left(X_{(i)}'X_{(i)}\right)^{-1}x_i\right]$，其估计为 $\hat{\sigma}^2_{(i)}\left[1 + x_i'\left(X_{(i)}'X_{(i)}\right)^{-1}x_i\right]$；

4.若 y_i 不是异常值，则 $E(y_i - \hat{y}_{(i)}) = 0$，在误差正态时，检验假设 $E(y_i - \hat{y}_{(i)}) = 0$ 的 t 检验统计量为

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x_i'\left(X_{(i)}'X_{(i)}\right)^{-1}x_i}},$$

给定水平 α，当 $|t_i| > t_{n-p-1}(\alpha/2)$ 时拒绝原假设，即认为 i 数据异常。

称 t_i为“学生化外残差”(因为σ^2的估计没有用到第 i个案例)。

定理 5.5.2：$t_i = \dfrac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$。

通常研究者对异常值无先验选择，若检验具有最大 t_i值的案例为异常值，则事实上进行了n次检验，对每个案例都进行了 t检验。

Box-Cox 变换

对观测得到的试验数据集(x_i', y_i)，$i = 1, \cdots, n$.若经回归诊断后发现不满足 GM 条件，就需要对数据采取“治疗”措施。数据变换是处理有问题数据的一种好的方法。Box 和 Cox(1964)对选择变换的问题给出了一个系统化的处理方法。实践证明，Box-Cox 变换对许多实际数据都是行之有效的，一般可明显地改善数据的正态性、方差齐性。

Box-Cox 变换是对回归响应变量作如下变换：

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}。$$

λ是待定的变换参数。Box 和 Cox 建议检验模型 $Y^{(\lambda)} = X\beta + e$，$Ee = 0$，$Var(e) = \sigma^2 I_n$，可用极大似然估计来确定$\lambda$。具体地，假设$Y^{(\lambda)} \sim N(X\beta, \sigma^2 I_n)$，似然函数为：

$$L(\beta, \sigma^2 | Y^{(\lambda)}) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(Y^{(\lambda)} - X\beta)'(Y^{(\lambda)} - X\beta)\right\}。$$

由于$\left|\dfrac{\partial Y^{(\lambda)}}{\partial Y}\right| = \prod_{i=1}^{n} y_i^{\lambda-1}$，故

$$L(\beta, \sigma^2, \lambda | Y) \propto \sigma^{-n} \prod_{i=1}^{n} y_i^{\lambda-1} \exp\left\{-\frac{1}{2\sigma^2}(Y^{(\lambda)} - X\beta)'(Y^{(\lambda)} - X\beta)\right\}。$$

先固定λ，β, σ^2的极大似然估计分别为

$$\hat{\beta}(\lambda) = (X'X)^{-1}X'Y^{(\lambda)},$$

$$\hat{\sigma}^2(\lambda) = \frac{Y^{(\lambda)'}(I_n - P_X)Y^{(\lambda)}}{n} = \frac{ESS(\lambda, Y^{(\lambda)})}{n}。$$

对应似然函数最大值为

$$L(\lambda)=\max_{\beta,\sigma^2} L(\beta,\sigma^2,\lambda|Y)$$

$$=\left(\frac{2\pi e}{n}\right)^{-\frac{n}{2}}\left[ESS(\lambda,Y^{(\lambda)})\right]^{-\frac{n}{2}}\prod_{i=1}^{n}y_i^{\lambda-1}。$$

对上式求最大值，最大值点 $\hat{\lambda}$作为 λ的极大似然估计。取对数似然，略去无关常数得

$$\ln L(\lambda)=-\frac{n}{2}\ln ESS(\lambda,Y^{(\lambda)})+\ln\prod_{i=1}^{n}y_i^{\lambda-1}。$$

令

$$z_i^{(\lambda)}=\begin{cases} y_i^{(\lambda)}\left(\prod_{i=1}^{n}y_i\right)^{\frac{1-\lambda}{n}}, & \lambda\neq 0 \\ (\ln y_i)\left(\prod_{i=1}^{n}y_i\right)^{\frac{1}{n}}, & \lambda=0 \end{cases},$$

$$Z^{(\lambda)}=(z_1^{(\lambda)},\cdots,z_n^{(\lambda)})',$$

则$\ln L(\lambda)=-\frac{n}{2}\ln ESS(\lambda,Z^{(\lambda)})$，故

$$\hat{\lambda}=Arg\min_{\lambda} ESS(\lambda,Z^{(\lambda)}),$$

其中$ESS(\lambda,Z^{(\lambda)})=Z^{(\lambda)'}(I_n-P_X)Z^{(\lambda)}$。最后得到$\beta,\sigma^2$的估计$\hat{\beta}(\hat{\lambda}),\hat{\sigma}^2(\hat{\lambda})$。

5.6 共线性、岭估计与主成分分析

矩阵X的列向量线性相关等价于方阵$X'X$为奇异矩阵，也等价于$X'X$有 0 特征值。一般称X的列向量近似线性相关，若$X'X$至少有一个很小(接近 0)的特征值。

考虑中心化标准化后的线性回归模型

$$y_i=\alpha+\beta_1\frac{x_{i1}-\bar{x}_1}{s_1}+\cdots+\beta_{p-1}\frac{x_{i,p-1}-\bar{x}_{p-1}}{s_{p-1}}+e_i,$$

$i=1,\cdots,n$，这里 $\bar{x}_j=\sum_{i=1}^{n}x_{ij}/n$，$s_j^2=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$，$1\le j\le p-1$.写成矩阵形式：$Y=\alpha\cdot E_n+X\beta+e$，$Ee=0$，$Var(e)=\sigma^2 I_n$，$rank(X_{n\times(p-1)})=p-1$。

常数项与回归系数的估计为 $\hat{\alpha}=\bar{y}$，$\hat{\beta}=(X'X)^{-1}X'Y$。若设计矩阵$X$的列向量近似线性相关，则称回归模型协变量之间是(近似共线性的(collinearity)，或者称设计矩阵X是共线性的(一般说来，设计矩阵精确共线性是偶然的)。

度量共线性的程度一般有两种。一种度量是 $X'X$ 有特征值$<\varepsilon$，此时称矩阵 X 为ε ill-defined。另外一个是从条件数(condition number)，矩阵 X 的条件数定义为

$cond(X)=\sqrt{\frac{\lambda_{\max}(X'X)}{\lambda_{\min}(X'X)}}$，条件数越大表明矩阵越病态，近似共线性程度越高。

设计矩阵共线性程度对回归系数的估计有着重要的影响。一般衡量估计$\hat{\beta}$的好坏用均方误差(mean squared error)$MSE(\hat{\beta})=E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$。由于

$$\begin{aligned}MSE(\hat{\beta})&=tr\left[E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right]\\&=tr\left[Var(\hat{\beta})\right]\\&=\sigma^2 tr(X'X)^{-1}\end{aligned},$$

设$\lambda_1\ge\cdots\ge\lambda_{p-1}>0$为$X'X$的特征值，则

$$MSE(\hat{\beta})=\sigma^2\sum_{i=1}^{p-1}\frac{1}{\lambda_i}。$$

若$X'X$有一个特征值非常小，则$MSE(\hat{\beta})$就会相当大，从均方误差来看，$\hat{\beta}$的 LS 估计就不是好的估计。

共线性产生的一般可能与数据收集有关，或者可能与回归协变量的选择有关。Hoerl 和 Kennard(1970)建议用

$$\hat{\beta}(k)=(X'X+kI_{p-1})^{-1}X'Y$$

作为β的估计(选择适当的常数$k>0$)，称为岭估计(ridge estimate)。

注：$\hat{\beta}(k)$是有偏的估计。

直观上看，当 X 呈病态时，$X'X$ 的特征值至少有一个非常接近与 0，故$X'X+kI_{p-1}$的特征值接近于 0 的程度将得到改善。

定理 5.6.1：$\exists k>0$使得

$$MSE(\hat{\beta}(k))<MSE(\hat{\beta})。$$

为证明上面的基本定理需要用到线性回归模型的典则形式(canonical form)。设$\lambda_1,\cdots,\lambda_{p-1}$为$X'X$的特征值，$\phi_1,\cdots,\phi_{p-1}$为对应的标准正交化特征向量，即$X'X=\Phi\Lambda\Phi'$，这里$\Lambda=diag(\lambda_1,\cdots,\lambda_{p-1})$，$\Phi=(\phi_1,\phi_2,\cdots\phi_{p-1})$。线性模型可以写成

$Y=\alpha\cdot E_n+Z\gamma+e$，$Ee=0$，$Var(e)=\sigma^2 I_n$，这里$Z=X\Phi$，$\gamma=\Phi'\beta$，此形式称为线性模型的典则形式。

此时参数估计为$\hat{\alpha}=\bar{y}$，$\hat{\gamma}=\Lambda^{-1}Z'Y$，$Var(\hat{\gamma})=\sigma^2\Lambda^{-1}$。典则形式回归系数$\gamma$的岭估计为

$$\begin{aligned}\hat{\gamma}(k)&=(Z'Z+kI_{p-1})^{-1}Z'Y\\&=(\Lambda+kI_{p-1})^{-1}Z'Y\end{aligned},$$

$$\hat{\beta}(k)=(X'X+kI_{p-1})^{-1}X'Y=\Phi\hat{\gamma}(k),$$

因此

$$MSE(\hat{\gamma})=MSE(\hat{\beta}),$$
$$MSE(\hat{\gamma}(k))=MSE(\hat{\beta}(k))。$$

定理 5.6.1 只需对典则形式证明即可。由于

$$MSE(\hat{\gamma}(k))=\sum_{i=1}^{p-1}\frac{k^2\gamma_i^2}{(\lambda_i+k)^2}+\sigma^2\sum_{i=1}^{p-1}\frac{\lambda_i}{(\lambda_i+k)^2},$$

$$\frac{\partial MSE(\hat{\gamma}(k))}{\partial k}=2\sum_{i=1}^{p-1}\frac{\lambda_i(k\gamma_i^2-\sigma^2)}{(\lambda_i+k)^3},$$

由于$\frac{\partial MSE(\hat{\gamma}(k))}{\partial k}\Big|_{k=0}<0$，$\exists k^*$使得$k\in[0,k^*)$，$\frac{\partial MSE(\hat{\gamma}(k))}{\partial k}<0$，$MSE(\hat{\gamma}(k))$为单调减函数，故$MSE(\hat{\gamma}(k))<MSE(\hat{\gamma}(0))=MSE(\hat{\gamma})$。

在实际应用中，岭参数 k的选择很重要。由于$MSE(\hat{\beta}(k))$依赖于未知参数β,σ^2，故 k不能从$\frac{\partial MSE(\hat{\gamma}(k))}{\partial k}=0$得到。Hoerl 和 Kennard 建议选择 k的估计为$\hat{k}=\frac{\hat{\sigma}^2}{\max\limits_{1\le i\le p-1}\hat{\gamma}_i^2}$。这个方法基于如下考虑，若对$i=1,\cdots,p-1$，$k\gamma_i^2-\sigma^2<0$，则一定有$\frac{\partial MSE(\hat{\beta}(k))}{\partial k}<0$。

当设计阵X存在多重共线性时（往往协变量维数过高），即有一些$X'X$的特征值很小时，一个解决多重共线性常用的方法是主成分分析(principal component analysis，PCA)。对线性模型中心化得到$Y=\alpha\cdot E_n+X\beta+e$，同上设$\lambda_1\ge\cdots\ge\lambda_{p-1}>0$为$X'X$的特征值，$\phi_1,\cdots,\phi_{p-1}$为对应的标准正交化特征向量，利用线性模型典则形式$Y=\alpha\cdot E_n+Z\gamma+e$，所谓主成分分析是指：选择$\lambda_1,\cdots,\lambda_q$使得

$\sum_{i=1}^{q}\lambda_i/\sum_{i=1}^{p-1}\lambda_i\ge c$，通常$c=80\%$，85%或者更大，由研究者自己选择；再选取$\lambda_i$所对应的$\phi_i$，$z_i=X\phi_i$，用新模型$Y=\alpha\cdot E_n+\tilde{Z}\tilde{\gamma}+e$

其中$\tilde{Z}_{n\times q}=X(\phi_1,\cdots,\phi_q)$，$\tilde{\gamma}=(\gamma_1,\cdots,\gamma_q)'$，得到LS 估计$\hat{\alpha}=\bar{y}$，$\hat{\tilde{\gamma}}=(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'Y$。其中$z_i$称为第$i$个主成分。当设计矩阵存在多重共线性时，适当选择保留主成分的个数可使主成分估计比原模型 LS 估计有较小的均方误差。主成分估计也是有偏估计。

主成分也等价于下面优化问题的解。

设X已中心化，第一个主成分$z_1=X\phi_1$，其中

$$\phi_1=Arg\max_{\|\phi\|=1}\phi'X'X\phi,$$

第i个主成分$z_i=X\phi_i$，其中

$$\phi_i=Arg\max_{\substack{\|\phi\|=1\\ \phi_l'X'X\phi=0,\ l=1,\cdots i-1.}}\phi'X'X\phi,$$

选择合适的$q\le p-1$个主成分，最后得到回归$\hat{y}^{PCR}=\bar{y}+\sum_{i=1}^{q}\hat{\gamma}_i z_i$。

5.7 偏最小二乘与逐步回归

当协变量维数较高时，设计矩阵往往容易存在共线性，不像上一节的主成分分析(PCA)只用了协变量的数据，并没有考虑响应变量，**偏最小二乘**(partial least squares，PLS)回归利用响应变量Y来构造偏最小二乘方向。

PLS 方法等价于下面优化问题的解。

设X，Y都已中心化，第一个 PLS 方向为$z_1 = X\phi_1$，其中

$$\phi_1 = Arg \max_{\|\phi\|=1} \phi' X'YY'X\phi,$$

第i个 PLS 方向为$z_i = X\phi_i$，其中

$$\phi_i = Arg \max_{\|\phi\|=1 \atop \phi' X'X\phi=0,\ l=1,\cdots i-1.} \phi' X'YY'X\phi,$$

选择合适的$q \le p-1$个 PLS 方向，最后得到回归$\hat{y}^{PLS} = \bar{y} + \sum_{i=1}^{q} \hat{\gamma}_i z_i$。

逐步回归

当协变量维数较高时，若通过所有可能协变量子集来确定最好的回归，例如 5.4 节中的方法，则在计算时间上可能不可行。**逐步回归**(stepwise regression)是解决此问题的一个常用方法。逐步回归一般可以采取向前选择(forward selection, FS)、向后削去(backward elimination, BE)或者交叉逐步(hybrid stepwise)进行。

向前选择 (forward selection, FS)

此方法从只有截距项开始，依次添加协变量。设当前已有k个协变量(含截距项)，回归系数估计为$\hat{\beta}$，残差平方和为$ESS(\hat{\beta})$，设新加一个协变量后回归系数估计为$\tilde{\beta}$，残差平方和为$ESS(\tilde{\beta})$，这样计算F值

$$F=\frac{ESS(\hat{\beta})-ESS(\tilde{\beta})}{ESS(\tilde{\beta})/(n-k-1)},$$

若有$F>F_{1,n-k-1}(\alpha)$，则具有最大F值所对应的协变量将添加进去；否则停止。

向后削去 (backward elimination)

此方法从所有协变量开始，依次剔除协变量。设当前已有$k+1$个协变量(含截距项)，回归系数估计为$\tilde{\beta}$，残差平方和为$ESS(\tilde{\beta})$，设剔除一个协变量后回归系数估计为$\hat{\beta}$，残差平方和为$ESS(\hat{\beta})$，这样计算F值

$$F=\frac{ESS(\hat{\beta})-ESS(\tilde{\beta})}{ESS(\tilde{\beta})/(n-k-1)},$$

若有$F<F_{1,n-k-1}(\alpha)$，则具有最小F值所对应的协变量将被剔除；否则停止。

交错逐步(hybrid stepwise)回归就是轮流使用添加和剔除的方式来选择变量。

5.8 Linear Regression with Regularization

设具有s个协变量的线性回归模型：

$$y_i=\beta_1 x_{i1}+\cdots+\beta_s x_{i,s}+e_i,\ i=1,2\cdots n$$

考虑一般形式地 penalized least-squares 方法，对给定非负罚函数(penalty function)$P(\beta)$，可以写成下面形式

$$\phi(\beta)=(Y-X\beta)'(Y-X\beta)+\lambda P(\beta)$$

其中$\beta=(\beta_1,\cdots\beta_s)'$为回归系数，$\lambda>0$称为 regularization parameter 或者 tuning parameter。$\hat{\beta}=Arg\min_{\beta}\phi(\beta)$。

如果$\lambda \to 0$，这样就得到通常最小二乘估计

$$\hat{\beta}_{LS} = Arg\min_{\beta}(Y - X\beta)'(Y - X\beta)。$$

一般常见罚函数取$P(\beta) = \sum_{j=1}^{s} p(\beta_j)$，其中$p(x) = p(|x|)$为偶函数。例如一般取$P_q(\beta) = \sum_{j=1}^{s} p_q(\beta_j) = \sum_{j=1}^{s} |\beta_j|^q$，相当于$p(x) = |x|^q$。

特别$q = 2$时，$p_2(x) = x^2$，将得到 ridge estimates

$$\hat{\beta}(\lambda) = (X'X + \lambda I_s)^{-1} X'Y。$$

求解

$$\min_{\beta} \phi(\beta) = \min_{\beta}(Y - X\beta)'(Y - X\beta) + \lambda P(\beta) \Leftrightarrow$$

对某个$t > 0$，在$P(\beta) \le t$约束条件下来求解$\min(Y - X\beta)'(Y - X\beta)$。

若当罚函数取$P_q(\beta)$时，定义相应的$\phi_q(\beta)$。取$p(x) = |x|^q$，称为L_q型罚函数。可见当$q > 1$时，$p_q(x) = |x|^q$，此时$\phi_q(\beta)$为光滑的凸函数；当$q = 1$时，$p_1(x) = |x|$，$\phi_q(\beta)$为凸函数，此时$\min_{\beta} \phi_q(\beta)$称为 Lasso(Tibshirani, 1996)。

当$q \ge 1$，可以用传统优化方法来求解$\min_{\beta} \phi_q(\beta)$。而当$0 < q < 1$时，$p_q(x) = |x|^q$，此时$\phi_q(\beta)$不是凸函数，求解较复杂。当$q = 0$时，此时$p_0(x) = I(x \ne 0)$，$P_q(\beta) = \sum_{j=1}^{s} I(\beta_j \ne 0)$，$\min_{\beta} \phi_q(\beta)$等价于$AIC$或 BIC 类型的变量选择准则。

注意：若模型含有常数项，通常是不被惩罚的。

注意到

$$(Y - X\beta)'(Y - X\beta) =$$

$$(Y - X\hat{\beta}_{LS})'(Y - X\hat{\beta}_{LS}) + (\beta - \hat{\beta}_{LS})' X'X (\beta - \hat{\beta}_{LS})$$

因此，

在$P_q(\beta) \le t$条件下，$\min(Y - X\beta)'(Y - X\beta) \Leftrightarrow$

$$\min_{\beta}(\beta - \hat{\beta}_{LS})' X'X (\beta - \hat{\beta}_{LS})$$

$$s.t. \quad \sum_{i=1}^{s} |\beta_i|^q \le t.$$

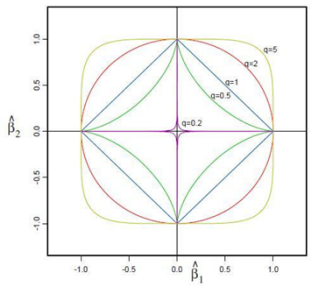
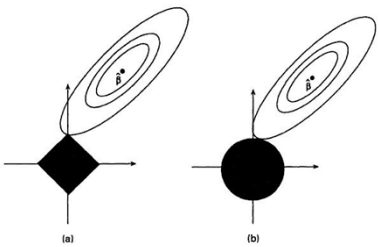


FIGURE *Two-dimensional contours of the symmetric penalty function $p_q(\beta) = |\beta_1|^q + |\beta_2|^q = 1$ for $q = 0.2, 0.5, 1, 2, 5$. The case $q = 1$ (blue diamond) yields the lasso and $q = 2$ (red circle) yields ridge regression.*

当$0 \le q \le 1$时，区域$\sum_{j=1}^{s}|\beta_j|^q \le t$有“角”(corners)，(与坐标轴的交点)，最小值通常都会在此处达到。这样所得的估计$\hat{\beta}$，其某些分量就是 0，从而表明所对应的协变量不在回归协变量中，同时达到变量选择和回归系数的估计。这种自动将某些回归系数估计为 0，从而降低模型复杂程度的估计方法，称为估计的稀疏性(sparsity)。下图解释了 Lasso($q=1$)具有稀疏性。



Estimation picture for (a) the lasso and (b) ridge regression

Fan & Li(2001)指出，一个好的罚函数选择应该使得估计具有以下三个性质：1)近似无偏性(approximately unbiased)（when the true unknown parameter is large to avoid unnecessary modeling bias）；2)稀疏性(sparsity)；3)连续性(continuity)(to avoid instability in prediction)。同时给出

1）满足近似无偏性的充分条件是罚函数满足$\lim_{|x|\to\infty} p'(|x|) = 0$；

2）满足稀疏性的一个充分条件是$\min_x |x| + p'(|x|) > 0$；

3）满足连续性的一个充分必要条件是$\min_x |x| + p'(|x|)$在$x = 0$处达到。

由 2)和 3)可知，同时满足稀疏性和连续性必须是罚函数$p(x)$在 0 点奇异，即在 0 处导数不存在。由此可见对于L_q型的罚函数$p_q(x)=|x|^q$，只有当$0\le q\le 1$，才有稀疏性；$q>1$不具有稀疏性。$q\ge 1$具有连续性，而$0\le q<1$不具有连续性。因此，$q=1$，即 Lasso 具有连续性，又具有稀疏性，而且还是凸函数。

Lasso 的缺点：

Lasso shrinks the estimates for the nonzero coeficients too heavily.

假设真实的模型所对应的回归系数为β_0，其中可以分成两个部分$(\beta_{10}',\beta_{20}')'$，其中$\beta_{20}=0$，即真实模型只包含$\beta_{10}$所对应的协变量。Fan & Li(2001)提出，一个好的变量选择程序得到的估计$\hat{\beta}=(\hat{\beta}_1',\hat{\beta}_2')'$在通常正则条件下应该具有如下两个性质，也称为 Oracle Properties：

1) (model selection consistency):

$$\lim_{n\to\infty} P(\hat{\beta}_2=0)=1;$$

2) (asymptotic normality):

$$\sqrt{n}(\hat{\beta}_1-\beta_{10})\xrightarrow{d.f.}N(0,\Sigma),$$

其中Σ是真实模型下的渐近方差。

在通常条件下，Lasso 不具有 Oracle 性质，只是在一个特殊条件下才有，见 Zhao and Yu (2006)。而对于$q<1$，其解具有 Oracle 性质，但不具有连续性，求解是非凸优化，较复杂。

Zou(2006)提出 Adaptive Lasso，对系数β给予不同的惩罚，对最小二乘估计$\hat{\beta}_{LS}$其绝对值大的，惩罚应该较小，即最小化

$$\min_{\beta}(Y-X\beta)'(Y-X\beta)+\lambda\sum_{j=1}^{s}w_j|\beta_j|$$

其中$w_j=1/\left|(\hat{\beta}_{LS})_j\right|^k, if\ (\hat{\beta}_{LS})_j\neq 0$，否则$w_j=0$。

第七章 线性混合效应模型

7.1 纵向数据与线性混合效应模型

在前面 1.6 节中虽然已经提到过纵向数据(Longitudinal data)以及线性混合效应模型(LMM)。这里，我们再回忆一下。一般纵向数据是指随时间演进而追踪测得的数据。更确切地说，在一项研究中，我们准备了一定数量K个个体，对每个个体随时间演进作了测量：对个体i在时刻$t_{i1}<t_{i2}<\cdots t_{in_i}$，测量得到$y_{i1},y_{i2},\cdots y_{in_i}$，$i=1,2,\cdots K$。

则$\left\{y_{ij}:1\le j\le n_i,1\le i\le K\right\}$就是纵向数据。对固定的$i$，$y_{i1},y_{i2},\cdots y_{in_i}$构成一个时间序列。因此，纵向数据也可以说由一批长度不等的时间序列构成。而通常的“时间序列分析”是指对一个较长的时间序列数据(例如某特定股市逐日的数据)来作统计分析。这一点把纵向数据与时间序列分别开来。

医学领域是纵向数据分析最主要的应用领域。因为一种药品或治疗方法的效果要在患者身上作时间追踪调查。

此外，社会经济领域也是重要的应用方面。对这方面有另外一个称呼：Panel data。Panel 是分组的意思，每个个体的量测数据成为一组(也许并不牵涉时间)。

纵向数据也是所谓“集团数据”(Clustered data)的一种。后者是指全部数据因某种共性而划分为一些小集团，例如一对父母所生的子女，同一地域或空间所采集的数据等。其共性都表现在看成同一个体上采集的数据。因此，处理纵向数据的统计方法不少也可用于集团数据。

关于纵向数据有一个基本假设：不同个体的测量是相互独立的，即$\left(y_{i1},y_{i2},\cdots y_{in_i}\right),1\le i\le K$是相互独立的。

令$Y_i=\left(y_{i1},y_{i2},\cdots y_{in_i}\right)'$(为$n_i\times 1$列向量)，从而对纵向数据的建模问题就归结为对$Y_i$的分布作假定的问题。一般，若$Y_i$为连续型变量，通常假定其为多维正态分布，此时问题简化为对其期望EY_i以及协方差$Cov(Y_i)$的假定。

由于$y_{i1}, y_{i2}, \cdots y_{in_i}$是同一个个体上的量测，它们之间往往不再独立而是有某种相关性，其形式与具体问题有关。这种相关性有时不易刻画，也是纵向数据统计分析的难点所在。

当Y_i取离散值，或虽取连续值但非多维正态分布时，要确切的假定其联合分布一般不可能。所幸的是，近三十年发展起来的一种叫“估计方程”的方法只要求对EY_i有正确假定就行。

我们采用由 Laird & Ware(1982)发展的混合效应模型来对纵向数据建模，设

$$Y_i = X_i\beta + Z_i b_i + e_i,\ \ i = 1,2,\cdots K,$$

其中Y_i为$n_i \times 1$第i个个体的响应变量，X_i为$n_i \times p$固定效应协变量设计矩阵，β为$p \times 1$的未知的总体的固定效应，Z_i为$n_i \times m$的随机效应设计矩阵，b_i为$m \times 1$的随机效应，e_i为$n_i \times 1$的测量误差。

由纵向数据分析的基本假设，假定$\{b_i, e_i, i = 1,2,\cdots K\}$为零均值相互独立的随机向量。一般来说假设$\{b_i, 1 \le i \le K\}$独立同分布，$Cov(b_i) = D_{m\times m} \ge 0$，且$b_i, e_i$之间也是独立的。这样我们有$Cov(Y_i) = Z_i D Z_i + R_i$，其中$R_i = Cov(e_i) > 0$为$n_i \times n_i$阶矩阵。

称线性混合效应是平衡的，若所有n_i相等且随机效应部分的设计矩阵Z_i也相等。

计$n = \sum_{i=1}^{K} n_i$为所有观测总数，如果写成单一向量的形式：

$$Y = X\beta + Zb + e,$$

这里

$$Y_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{pmatrix},\ X_{n\times p} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{pmatrix},\ Z = \begin{pmatrix} Z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_K \end{pmatrix}$$

为$n \times (mK)$阶矩阵，

$$b_{(mK)\times 1}=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{pmatrix},\quad e_{n\times 1}=\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_K \end{pmatrix},$$

此时$Cov(b)=I_K\otimes D$。

注：⊗为矩阵的 Kronecker product，两矩阵$A=(a_{ij})_{m\times n}$，$B=(b_{ij})_{p\times q}$），则

$$A\otimes B=\begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

为$(mp)\times(nq)$阶矩阵。

记 $\eta=\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_K \end{pmatrix}=\begin{pmatrix} Z_1b_1+e_1 \\ Z_2b_2+e_2 \\ \vdots \\ Z_Kb_K+e_K \end{pmatrix}$，则有 $E\eta=0$，

$$Cov(\eta)=\begin{pmatrix} Z_1DZ_1'+R_1 & 0 & \cdots & 0 \\ 0 & Z_2DZ_2'+R_2 & & \vdots \\ \vdots & \vdots & & 0 \\ 0 & 0 & \cdots & Z_KDZ_K'+R_K \end{pmatrix}$$

得到边际模型(Marginal model)

$$Y=X\beta+\eta。$$

一般情形下，作为量测误差e，通常可以假定$Cov(e)=\sigma^2 I_n$(以下不特别声明都作此假定)，这意味着$R_i=\sigma^2 I_{n_i}$。若令$\tilde{D}=\dfrac{D}{\sigma^2}$，则

$$Cov(\eta)=\sigma^2\begin{pmatrix} Z_1\tilde{D}Z_1'+I_{n_1} & 0 & \cdots & 0 \\ 0 & Z_2\tilde{D}Z_2'+I_{n_2} & & \vdots \\ \vdots & \vdots & & 0 \\ 0 & 0 & \cdots & Z_K\tilde{D}Z_K'+I_{n_K} \end{pmatrix}$$

此时，对于线性混合效应模型来说，感兴趣的未知参数为$\theta=\left(\beta',\sigma^2,\left(vech(\tilde{D})\right)'\right)'$。

注：这里$vech(\tilde{D})$为$m\times m$阶对称矩阵中独立变化的$\frac{m(m+1)}{2}$个元素构成的列向量。

从而，感兴趣的未知参数θ的维数为$p+1+\frac{m(m+1)}{2}$。

7.2 极大似然估计

本节我们先来研究模型未知参数的估计。假定$e_i\sim N(0,\sigma^2 I_{n_i})$，$b_i\sim N(0,D)$，此时$Y_i\sim N\left(X_i\beta,\sigma^2(I_{n_i}+Z_i\tilde{D}Z_i')\right)$，$i=1,2,\cdots K$。略去无关常数项，对数似然为

$$l(\theta)=-\frac{1}{2}\left\{n\ln\sigma^2+\sum_{i=1}^{K}\ln\left|I_{n_i}+Z_i\tilde{D}Z_i'\right|+\frac{\sum_{i=1}^{K}(Y_i-X_i\beta)'\left(I_{n_i}+Z_i\tilde{D}Z'\right)^{-1}(Y_i-X_i\beta)}{\sigma^2}\right\}$$

这里θ由7.1节定义，记$\Theta=\left\{\theta\middle|\beta\in R^p,\sigma^2>0,\tilde{D}_{m\times m}\geq 0(\text{非负定})\right\}$，则参数极大似然估计(MLE)定义为

$$\hat{\theta}=Arg\max_{\theta\in\Theta}l(\theta)。$$

注：参数空间Θ是凸集。

为简化记号，以下记$V_i=V_i(\tilde{D})=I_{n_i}+Z_i\tilde{D}Z_i'$，$e_i(\beta)=Y_i-X_i'\beta$，则

$$l(\theta)=-\frac{1}{2}\left\{n\ln\sigma^2+\sum_{i=1}^{K}\left[\ln|V_i|+\frac{e_i(\beta)'V_i^{-1}e_i(\beta)}{\sigma^2}\right]\right\}。$$

由于

$$\begin{aligned}V_i^{-1}&=\left(I_{n_i}+Z_i\tilde{D}Z_i'\right)^{-1}=I_{n_i}-Z_i\left(I_m+\tilde{D}Z_i'Z_i\right)^{-1}\tilde{D}Z_i'\\&=I_{n_i}-Z_i\tilde{D}\left(I_m+Z_i'Z_i\tilde{D}\right)^{-1}Z_i'\\&=I_{n_i}-Z_i\left(\tilde{D}^{-1}+Z_i'Z_i\right)^{-1}Z_i'(\text{若}\tilde{D}^{-1}\text{存在})\\|V_i|&=\left|I_{n_i}+Z_i\tilde{D}Z'\right|=\left|I_m+\tilde{D}Z_i'Z_i\right|,\end{aligned}$$

这样，在对数似然函数中涉及到$n_i\times n_i$阶矩阵求逆就转化为$m\times m$阶矩阵的求逆。此即为dimension-reduction 技巧。

若假设$\tilde{D}^{-1}$存在，则由于

$\ln|V_i| = \ln|I_m + \tilde{D}Z_i'Z_i| = \ln|\tilde{D}^{-1} + Z_i'Z_i| - \ln|\tilde{D}^{-1}|$，

将$\tilde{D}^{-1}$看成未知参数，此时对数似然为$\beta, \sigma^2, \tilde{D}^{-1}$的函数。

此外，还可以采用 profile log-likelihood function(子集对数似然函数)技巧来求MLE。由于在给定β和$\tilde{D}$的条件下，σ^2的MLE 为

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{K} e_i'(\beta)V_i^{-1}e_i(\beta),$$

将其带入对数似然函数中削去σ^2，从而得到只基于β和$\tilde{D}$的函数，称为β和$\tilde{D}$的 profile log-likelihood function，略去无关常数，记为$l_{profile}(\beta, \tilde{D})$，则

$$l_{profile}(\beta, \tilde{D}) = -\frac{1}{2}\left\{n\ln\sum_{i=1}^{K} e_i'(\beta)V_i^{-1}e_i(\beta) + \sum_{i=1}^{K}\ln|V_i|\right\}。$$

进一步，给定$\tilde{D}$，即V_i已知的条件下，β的极大似然估计(这里即广义最小二乘估计)为

$$\hat{\beta} = Arg\min_{\beta}\sum_{i=1}^{K} e_i'(\beta)V_i^{-1}e_i(\beta)。$$

令$h(\tilde{D}) = \min_{\beta}\sum_{i=1}^{K} e_i'(\beta)V_i^{-1}e_i(\beta)$，则可以得到基于$\tilde{D}$的函数，称为$\tilde{D}$的 profile log-likelihood function，略去无关常数，记为$l_{profile}(\tilde{D})$，则

$$l_{profile}(\tilde{D}) = -\frac{1}{2}\left\{n\ln h(\tilde{D}) + \sum_{i=1}^{K}\ln|V_i(\tilde{D})|\right\}。$$

从而得到$\tilde{D}$的极大似然估计

$$\hat{\tilde{D}} = Arg\max_{\tilde{D}\geq 0} l_{profile}(\tilde{D})。$$

7.3 限制极大似然估计

对于正态分布来说，一般方差的极大似然估计是有偏的。例如对标准正态线性模型$Y \sim N(X\beta, \sigma^2)$，令$ESS$为残差平方和，则$\sigma^2$的极大似然估计为$\frac{ESS}{n}$是有偏的，会低估$\sigma^2$。其无偏估计为$\frac{ESS}{n-p}$，其中$p$为未知参数$\beta$的维数。其原因在于$Y$的作用有一部分用于估计$\beta$了。因此用于估计$\sigma^2$的自由度不是$n$，而是比$n$小，要调整。对于上一节正态分布假定

的线性混合效应模型，要估计协方差阵σ^2, D或者$\sigma^2, \tilde{D}$(这里$\tilde{D} = \frac{D}{\sigma^2}$)，不止一个参数。因此不是单由调整自由度就能解决。

先看一般线性模型限制极大似然估计(Restricted maximum likelihood estimation, RMLE)的做法，然后将其用于线性混合效应模型。设$Y \sim N(X\beta, V)$，这里$X_{n\times p}$列满秩矩阵，$V_{n\times n}$为协方差矩阵依赖于某些未知参数。

考虑Y的两个线性成分：

$$T_{p\times 1} = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

$$U_{(n-p)\times 1} = BY$$

其中B为$(n-p)\times n$阶矩阵满足$B'B = I_n - P_X$，$BB' = I_{n-p}$。

注：这里T即为V已知时β的广义最小二乘估计。由于$I_n - P_X$是幂等矩阵，因此存在$n\times n$正交阵Q使得$I_n - P_X = Q\begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix}Q'$。记

$Q = (Q_1, Q_2)$，其中Q_1为$n\times(n-p)$，则有$Q_1Q_1' = I_n - P_X$。由于$QQ' = I_n$，从而有$Q_1'Q_1 = I_{n-p}$。因此只要取$B = Q_1'$即可。且

$$BX = (BB')BX = B(I_n - P_X)X = 0。$$

这样，得到$ET = \beta$，$EU = BX\beta = 0$。由此看出，T这部分用于估计β，于估计协方差阵无益。而U则不然，其变异纯由协方差阵而来，不涉及到β。这部分正是估计协方差时该用上的。

以上就是 RMLE 的想法。为实施这一想法，须求出U的分布。为此，先计算$\begin{pmatrix} T \\ U \end{pmatrix}$的联合分布。由于

$$\begin{pmatrix} T \\ U \end{pmatrix} = \begin{pmatrix} (X'V^{-1}X)^{-1}X'V^{-1} \\ B \end{pmatrix} Y \equiv CY,$$

这里$C = \begin{pmatrix} (X'V^{-1}X)^{-1}X'V^{-1} \\ B \end{pmatrix}$，从而

。

$\begin{pmatrix} T \\ U \end{pmatrix}$的联合密度=

Y的密度×变换的Jacobi行列式的绝对值

=Y的密度×$\left|C^{-1}\right|$的绝对值

由于

$$Cov(T,U)=(X'V^{-1}X)^{-1}X'V^{-1}Cov(Y)B'$$
$$=(X'V^{-1}X)^{-1}X'B'=0,$$

故T,U独立。

$$U\text{的密度}=\frac{Y\text{的密度}\times\left|C^{-1}\right|\text{的绝对值}}{T\text{的密度}}$$

而$Y\sim N(X\beta,V)$，$T\sim N(\beta,(X'V^{-1}X)^{-1})$，

Y的密度为

$$f(Y)=(2\pi)^{-\frac{n}{2}}|V|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(Y-X\beta)'V^{-1}(Y-X\beta)\right],$$

T的密度为

$$g(T)=(2\pi)^{-\frac{p}{2}}\left|X'V^{-1}X\right|^{\frac{1}{2}}\exp\left[-\frac{1}{2}(T-\beta)'(X'V^{-1}X)(T-\beta)\right]$$

而

$$(Y-X\beta)'V^{-1}(Y-X\beta)$$
$$=[Y-XT+X(T-\beta)]'V^{-1}[Y-XT+X(T-\beta)],$$
$$=(Y-XT)'V^{-1}(Y-XT)+(T-\beta)'(X'V^{-1}X)(T-\beta)$$

从而

U的密度=$\left|C^{-1}\right|$的绝对值×

$$(2\pi)^{-\frac{n-p}{2}}|V|^{-\frac{1}{2}}\left|X'V^{-1}X\right|^{\frac{1}{2}}\exp\left[-\frac{1}{2}(Y-XT)'V^{-1}(Y-XT)\right]$$

由于$\left|C\right|^2=\left|X'X\right|^{-1}$，与$V$无关。

注：由C的定义，

$C'C=V^{-1}X(X'V^{-1}X)^{-2}X'V^{-1}+I_n-P_X$。

由于$X_{n\times p}$列满秩，其 SVD(奇异值)分解为

$X=H_{n\times n}\begin{pmatrix}\Lambda_p\\0\end{pmatrix}O_{p\times p}$，其中$H,O$为正交矩阵。

令$H'V^{-1}H=\begin{pmatrix}V_1 & V_2\\ V_2' & V_3\end{pmatrix}$，则

$$H'C'CH=\begin{pmatrix}V_1\\V_2'\end{pmatrix}V_1^{-1}\Lambda_p^{-2}V_1^{-1}\begin{pmatrix}V_1 & V_2\end{pmatrix}+\begin{pmatrix}0 & 0\\0 & I_{n-p}\end{pmatrix}。$$

$$\begin{pmatrix} I_p & 0 \\ -V_2'V_1^{-1} & I_{n-p} \end{pmatrix} H'C'CH$$

$$= \begin{pmatrix} V_1 \\ 0 \end{pmatrix} V_1^{-1}\Lambda_p^{-2}V_1^{-1}\begin{pmatrix} V_1 & V_2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & I_{n-p} \end{pmatrix}$$

$$= \begin{pmatrix} \Lambda_p^{-2} & \Lambda_p^{-2}V_1^{-1}V_2 \\ 0 & I_{n-p} \end{pmatrix},$$

两边取行列式，得$|C|^2 = |\Lambda_p|^{-2} = |X'X|^{-1}$，

因此$|C^{-1}|$与V无关。

RMLE 就是从U出发，用极大似然方法来估计V。略去与V无关的量，取对数得到

$$l_R(V) = -\frac{1}{2}\left[\ln|V| + \ln|X'V^{-1}X| + (Y - XT)'V^{-1}(Y - XT)\right]$$

(注意：T中含有V)，从而得到V的估计

$$\hat{V} = Arg\max_{V>0} l(V),$$

代入T中，得到β的 RMLE 估计

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y。$$

关于$l_R(V)$的推导还可以从另外一个角度。将β看成随机，$Y|\beta \sim N(X\beta, V)$。即

$$f(Y|\beta) = (2\pi)^{-\frac{n}{2}}|V|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta)\right]$$

取β的先验分布为广义先验，即$\pi(\beta) \propto 1$，则由 Bayes 公式，Y的边际分布为

$$(2\pi)^{-\frac{n-p}{2}}|V|^{-\frac{1}{2}}|X'V^{-1}X|^{\frac{1}{2}}\exp\left[-\frac{1}{2}(Y - XT)'V^{-1}(Y - XT)\right]$$

略去无关常数，取对数后与$l_R(V)$一致。

将 RML 与标准的似然函数比较，发现多了一个因子$|X'V^{-1}X|^{-\frac{1}{2}}$，对数似然函数多了一项$-\frac{1}{2}\ln|X'V^{-1}X|$。当$p$相对于$n$很小时，这一项影响不大，此时 RMLE 与 MLE 很接近。当$\frac{p}{n}$不接近 0 时，经验表明使用 RML 会使估计会得到改善。

以上是一般线性模型 RMLE 的做法，对线性混合效应模型来说，其对数限制似然函数为

$$l_R(\theta)=-\frac{1}{2}\left\{(n-p)\ln\sigma^2+\sum_{i=1}^{K}\ln\left|X_i'V_i^{-1}X_i\right|\right.$$
$$\left.+\sum_{i=1}^{K}\left[\ln|V_i|+\frac{e_i(\beta)'V_i^{-1}e_i(\beta)}{\sigma^2}\right]\right\}$$

相对于标准的线性混合效应模型对数似然函数作了调整，增加了两项

$$\frac{p}{2}\ln\sigma^2-\frac{1}{2}\sum_{i=1}^{K}\ln\left|X_i'V_i^{-1}X_i\right|$$

第二项就是刚才说的 RML 与标准的 ML 函数区别。第一项是对σ^2的因子也做了调整。此时在$\theta\in\Theta$上极大化$l_r(\theta)$得到 RMLE 估计。

7.4 Balanced random-coefficient model

本节考虑一类特殊的线性混合效应模型(LMM)。考虑第i个个体的回归模型

$$Y_i=X_i\gamma+e_i,\ 1\le i\le K$$

回归系数不能看成固定效应，而看成随机的

$$Y_i=X_i\gamma_i+e_i,\ 1\le i\le K$$

$\{\gamma_i,1\le i\le K\}$ 独立同分布，由于$\gamma_i=E\gamma_i+(\gamma_i-E\gamma_i)$，令$\beta=E\gamma_i$，$b_i=\gamma_i-E\gamma_i$，这样化成标准的 LMM

$$Y_i=X_i\beta+X_ib_i+e_i,\quad 1\le i\le K$$

此模型称为 random-coefficient model，是一类特殊的线性混合效应模型，其$Z_i=X_i$。再特殊一点，如果是平衡的，意味着所有$n_i\equiv s$且$Z_i\equiv Z$，$1\le i\le K$。此模型，MLE 有显示表达。此时，模型为

$$Y_i=Z\beta+Zb_i+e_i,1\le i\le K,$$

其中$Z_{s\times p}$，且$rank(Z)=p<s$。令$\eta_i=Zb_i+e_i$，则$E\eta_i=0$，$Cov(\eta_i)=\sigma^2(I_s+Z\tilde{D}Z')\equiv\sigma^2V$。

对此模型，β的最小二乘估计

$$\hat{\beta}_{OLS}=(Z'Z)^{-1}Z'\bar{Y},$$

这里$\bar{Y}_{s\times 1}=\frac{1}{K}\sum_{i=1}^{K}Y_i$。

当$\tilde{D}$已知时，可得β的广义最小二乘估计$\hat{\beta}_{GLS}$，定义为

$$\hat{\beta}_{GLS}=Arg\min_{\beta}\sum_{i=1}^{K}(Y_i-Z\beta)'V^{-1}(Y_i-Z\beta)。$$

注：通常$\hat{\beta}_{GLS}$并不是β的估计，因为含有未知的$\tilde{D}$。

定理 7.4.1：对此模型有$\hat{\beta}_{GLS}=\hat{\beta}_{OLS}$。

证明：由于

$$\sum_{i=1}^{K}(Y_i-Z\beta)'V^{-1}(Y_i-Z\beta)=\sum_{i=1}^{K}(Y_i-\bar{Y})'V^{-1}(Y_i-\bar{Y})+K(\bar{Y}-Z\beta)'V^{-1}(\bar{Y}-Z\beta)。$$

从而

$$\hat{\beta}_{GLS}=Arg\min_{\beta}\sum_{i=1}^{K}(Y_i-Z\beta)'V^{-1}(Y_i-Z\beta)$$
$$=Arg\min_{\beta}(\bar{Y}-Z\beta)'V^{-1}(\bar{Y}-Z\beta)$$

又由于$V^{-1}=\left(I_s+Z\tilde{D}Z'\right)^{-1}=I_s-Z\left(I_m+\tilde{D}Z'Z\right)^{-1}\tilde{D}Z'$及$(\bar{Y}-Z\hat{\beta}_{OLS})'Z=0$，有

$$(\bar{Y}-Z\beta)'V^{-1}(\bar{Y}-Z\beta)=(\bar{Y}-Z\hat{\beta}_{OLS})'V^{-1}(\bar{Y}-Z\hat{\beta}_{OLS})+(Z\hat{\beta}_{OLS}-Z\beta)'V^{-1}(Z\hat{\beta}_{OLS}-Z\beta)。$$

因此，当$\beta=\hat{\beta}_{OLS}$时达到最小值。

由此定理，对此模型有$\hat{\beta}_{GLS}=\hat{\beta}_{OLS}=\hat{\beta}_{ML}$，这里$\hat{\beta}_{ML}$表示$\beta$的极大似然估计。

定理 7.4.2：对此模型，其方差参数的极大似然(ML)估计及限制极大似然(RML)估计为

$$\hat{\sigma}^2_{ML}=\hat{\sigma}^2_{RML}=\frac{1}{K(s-p)}\sum_{i=1}^{K}Y_i'(I_s-Z(Z'Z)^{-1}Z')Y_i',$$

$$\tilde{D}_{ML}=\frac{1}{K\hat{\sigma}^2_{ML}}(Z'Z)^{-1}Z'\hat{E}\hat{E}'Z(Z'Z)^{-1}-(Z'Z)^{-1},$$

$$\tilde{D}_{RML}=\frac{1}{(K-1)\hat{\sigma}^2_{ML}}(Z'Z)^{-1}Z'\hat{E}\hat{E}'Z(Z'Z)^{-1}-(Z'Z)^{-1},$$

这里$\hat{E}\hat{E}'$为$s\times s$矩阵

$$\hat{E}\hat{E}'=\sum_{i=1}^{K}(Y_i-Z\hat{\beta}_{ML})(Y_i-Z\hat{\beta}_{ML})'。$$

7.5 LM Model with random intercepts

本节考虑另一类特殊的线性混合效应模型(LMM)，截距项具有随机效应，具体为：

$y_{ij} = a_i + w_{ij}'\gamma + e_{ij}$， $j = 1,2\cdots n_i, i = 1,2\cdots K$，

其中$\{a_i, 1 \le i \le K\}$独立同分布，为 random intercepts。由于 $a_i = Ea_i + (a_i - Ea_i)$，令$\alpha = Ea_i$，$b_i = a_i - Ea_i$，则$a_i = \alpha + b_i$。则模型为

$y_{ij} = a + w_{ij}'\gamma + b_i + e_{ij}$， $j = 1,2\cdots n_i, i = 1,2\cdots K$

这里随机效应b_i是一维的，假设$b_i \sim N(0,\sigma_b^2)$，由于通常假设$e_{ij} \sim N(0,\sigma^2)$，令$d = \frac{\sigma_b^2}{\sigma^2} \ge 0$，则$b_i \sim N(0,\sigma^2 d)$。写成向量形式

$$Y_i = X_i\beta + E_{n_i}b_i + e_i，\ 1 \le i \le K，$$

其中X_i为$n_i \times p$阶满秩矩阵，其第j行$x_{ij}' = (1, w_{ij}')$，$j = 1,2\cdots n_i$；$\beta = (\alpha, \gamma')'$为$p \times 1$固定效应，$E_{n_i}$为$n_i \times 1$元素全为 1 的向量。

注：相对于标准形式$m = 1, Z_i = E_{n_i}, \tilde{D} = d$。

由 7.2 节，可以得到基于β, d的 profile 对数似然

$$l_{profile}(\beta,d) = -\frac{1}{2}\left\{ n\ln\sum_{i=1}^{K} e_i'(\beta)(I_{n_i} + d\cdot E_{n_i}E_{n_i}')^{-1}e_i(\beta) + \sum_{i=1}^{K}\ln\left|I_{n_i} + d\cdot E_{n_i}E_{n_i}'\right|\right\},$$

这里$e_i(\beta) = Y_i - X_i\beta$为$n_i \times 1$向量。

由于

$$\left|I_{n_i} + d\cdot E_{n_i}E_{n_i}'\right| = 1 + n_i d，$$

$$\left(I_{n_i} + d\cdot E_{n_i}E_{n_i}'\right)^{-1} = I_{n_i} - \frac{d}{1+n_i d}E_{n_i}E_{n_i}'，$$

从而

$$l_{profile}(\beta,d) = -\frac{1}{2}\left\{ n\ln\left[S(\beta) - \sum_{i=1}^{K}\frac{n_i^2 d h_i^2(\beta)}{1+n_i d}\right] + \sum_{i=1}^{K}\ln(1+n_i d)\right\}$$

这里

$$S(\beta) = \sum_{i=1}^{K}\left\|Y_i - X_i\beta\right\|^2，$$

$$h_i(\beta) = \frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - x_{ij}'\beta) \equiv \bar{y}_i - \bar{x}_i'\beta。$$

其限制对数似然为

$$l_r(\beta,d)=-\frac{1}{2}\left\{(n-p)\ln\left[S(\beta)-\sum_{i=1}^{K}\frac{n_i^2 d h_i^2(\beta)}{1+n_i d}\right]+\sum_{i=1}^{K}\ln(1+n_i d)+\ln\left|\sum_{i=1}^{K}\left(X_i'X_i-\frac{n_i^2 d}{1+n_i d}\bar{x}_i\bar{x}_i'\right)\right|\right\}。$$

对于平衡观测(即$n_i\equiv s$)，profile 对数似然有简单表达

$$l_{profile}(\beta,d)=-\frac{1}{2}\left\{Ks\ln\left[S(\beta)-\frac{s^2 d}{1+sd}\sum_{i=1}^{K}h_i^2(\beta)\right]+K\ln(1+sd)\right\}$$

由$\frac{\partial l_{profile}(\beta,d)}{\partial d}=0$，得到$d$的 MLE(在给定$\beta$条件下)

$$d=\frac{s^2\sum_{i=1}^{K}h_i^2(\beta)-S(\beta)}{s\left[S(\beta)-s\sum_{i=1}^{K}h_i^2(\beta)\right]}$$

假设对平衡数据，这意味着$n_i\equiv s$， $X_i=X$，此时截距项具有随机效应的模型参数极大似然(ML)估计及限制极大似然(RML)估计有显示表示。

引理 7.5.1：记$X=\begin{pmatrix}E_s & U\end{pmatrix}$为$s\times p$阶列满秩矩阵，$\bar{x}=\frac{X'E_s}{s}$为$p\times 1$平均值向量，则有

1.$(X'X)^{-1}X'E_s=\left(1,0_{1\times(p-1)}\right)'$；

2.$X(X'X)^{-1}X'E_s=E_s$； 3.$\bar{x}'(X'X)^{-1}\bar{x}=\frac{1}{s}$。

对此模型，β的最小二乘估计

$$\hat{\beta}_{OLS}=(X'X)^{-1}X'\bar{Y}，$$

这里$\bar{Y}_{s\times 1}=\frac{1}{K}\sum_{i=1}^{K}Y_i$。

当d已知时，可得β的广义最小二乘估计$\hat{\beta}_{GLS}$，定义为

$$\hat{\beta}_{GLS}=Arg\min_{\beta}\sum_{i=1}^{K}(Y_i-X\beta)'\left(I_s+d\cdot E_sE_s'\right)^{-1}(Y_i-X\beta)。$$

注：通常$\hat{\beta}_{GLS}$并不是β的估计，因为含有未知的d。

定理 7.5.1：对此模型有$\hat{\beta}_{GLS}=\hat{\beta}_{OLS}$。

证明：由于

$$\begin{aligned}&\sum_{i=1}^{K}(Y_i-X\beta)'\left(I_s+d\cdot E_sE_s'\right)^{-1}(Y_i-X\beta)\\&=\sum_{i=1}^{K}(Y_i-\bar{Y})'\left(I_s+d\cdot E_sE_s'\right)^{-1}(Y_i-\bar{Y})\quad 。\\&+K(\bar{Y}-X\beta)'\left(I_s+d\cdot E_sE_s'\right)^{-1}(\bar{Y}-X\beta)\end{aligned}$$

从而

$$\begin{aligned}\hat{\beta}_{GLS}&=Arg\min_{\beta}\sum_{i=1}^{K}(Y_i-Z\beta)'V^{-1}(Y_i-Z\beta)\\&=Arg\min_{\beta}(\bar{Y}-Z\beta)'V^{-1}(\bar{Y}-Z\beta)\end{aligned}$$

由于$(\bar{Y}-X\hat{\beta}_{OLS})'X=0$，$(\bar{Y}-X\hat{\beta}_{OLS})'E_s=0$，

$$\begin{aligned}&(\bar{Y}-X\beta)'\left(I_s-\frac{d}{1+sd}E_sE_s'\right)(\bar{Y}-X\beta)\\&=(\bar{Y}-X\hat{\beta}_{OLS})'\left(I_s-\frac{d}{1+sd}E_sE_s'\right)(\bar{Y}-X\hat{\beta}_{OLS})\quad ,\\&+(X\beta-X\hat{\beta}_{OLS})'\left(I_s-\frac{d}{1+sd}E_sE_s'\right)(X\beta-X\hat{\beta}_{OLS})\end{aligned}$$

因此，当$\beta=\hat{\beta}_{OLS}$时达到最小值。

由此定理，对此模型有$\hat{\beta}_{GLS}=\hat{\beta}_{OLS}=\hat{\beta}_{ML}$，这里$\hat{\beta}_{ML}$表示$\beta$的极大似然估计。

前面已经推导过，在给定β,d时σ^2的极大似然(ML)估计满足

$$\begin{aligned}\sigma^2&=\frac{1}{Ks}\sum_{i=1}^{K}(Y_i-X\beta)'\left(I_s+d\cdot E_sE_s'\right)^{-1}(Y_i-X\beta)\\&=\frac{1}{Ks}\left[S(\beta)-\frac{s^2d}{1+sd}\sum_{i=1}^{K}h_i^2(\beta)\right]\end{aligned}$$

由于给定β，d的极大似然(ML)估计满足

$$d=\frac{s^2\sum_{i=1}^{K}h_i^2(\beta)-S(\beta)}{s\left[S(\beta)-s\sum_{i=1}^{K}h_i^2(\beta)\right]}=\frac{(s-1)\sum_{i=1}^{K}h_i^2(\beta)}{\left[S(\beta)-s\sum_{i=1}^{K}h_i^2(\beta)\right]}-\frac{1}{s}$$

将d代入，再将β用$\hat{\beta}_{ML}=\hat{\beta}_{OLS}$代入，得到

$$\hat{\sigma}_{ML}^2=\frac{S(\hat{\beta}_{ML})-s\sum_{i=1}^{K}h_i^2(\hat{\beta}_{ML})}{K(s-1)},$$

$$\hat{d}_{ML}=\frac{(s-1)\sum_{i=1}^{K}h_i^2(\hat{\beta}_{ML})}{\left[S(\hat{\beta}_{ML})-s\sum_{i=1}^{K}h_i^2(\hat{\beta}_{ML})\right]}-\frac{1}{s}=\frac{\sum_{i=1}^{K}h_i^2(\hat{\beta}_{ML})}{K\hat{\sigma}_{ML}^2}-\frac{1}{s}。$$

注：此特殊情形$h_i(\hat{\beta}_{ML})=\bar{y}_i-\bar{\bar{y}}$，这里$\bar{y}_i=\frac{E_s'Y_i}{s}$，$\bar{\bar{y}}=\frac{E_s'\bar{Y}}{s}$。

下面考察其限制极大似然(RML)估计。在此特殊情形下，$X_i \equiv X$

$$\begin{aligned}
&\left|\sum_{i=1}^{K}\left(X_i'X_i - \frac{s^2 d}{1+sd}\bar{x}_i\bar{x}_i'\right)\right| \\
&= \left|K(X'X - \frac{d}{1+sd}X'E_sE_s'X)\right| \\
&= \left|K(X'X)\right|\left|I_s - \frac{d}{1+sd}(X'X)^{-1}X'E_sE_s'X\right| \\
&= \left|K(X'X)\right|\left|1 - \frac{d}{1+sd}E_s'X(X'X)^{-1}X'E_s\right| \\
&= \left|K(X'X)\right|(1+sd)^{-1}
\end{aligned}$$

其限制对数似然(略去无关常数)也有简单表示

$$l_r(\beta,d) = -\frac{1}{2}\left\{(Ks-p)\ln\left[S(\beta) - \frac{s^2 d}{1+sd}\sum_{i=1}^{K}h_i^2(\beta)\right]\right.。$$
$$\left. + K\ln(1+sd) - \ln(1+sd)\right\}$$

由$\frac{\partial l_r(\beta,d)}{\partial d} = 0$，得到$d$的 RMLE(在给定$\beta$条件下)

$$d = \frac{[K(s-1)-p+1]\sum_{i=1}^{K}h_i^2(\beta)}{(K-1)\left[S(\beta) - s\sum_{i=1}^{K}h_i^2(\beta)\right]} - \frac{1}{s}。$$

由定理 7.5.1，对此模型有$\hat{\beta}_{GLS} = \hat{\beta}_{OLS} = \hat{\beta}_{ML} = \hat{\beta}_{RML}$，这里$\hat{\beta}_{RML}$表示$\beta$的限制极大似然估计。类似前面的推导，可以得到此时$\sigma^2$和$d$的限制极大似然(RML)估计如下：

$$\hat{\sigma}^2_{RML} = \frac{S(\hat{\beta}_{RML}) - s\sum_{i=1}^{K}h_i^2(\hat{\beta}_{RML})}{K(s-1)-p+1},$$

$$\begin{aligned}
\hat{d}_{RML} &= \frac{[K(s-1)-p+1]\sum_{i=1}^{K}h_i^2(\hat{\beta}_{RML})}{(K-1)\left[S(\hat{\beta}_{RML}) - s\sum_{i=1}^{K}h_i^2(\hat{\beta}_{RML})\right]} - \frac{1}{s} \\
&= \frac{\sum_{i=1}^{K}h_i^2(\hat{\beta}_{RML})}{(K-1)\hat{\sigma}^2_{RML}} - \frac{1}{s}
\end{aligned}$$

。

7.6 MINQUE for variance parameters

前面提到的线性混合效应模型(LMM)的 ML 估计及 RML 估计，都要假定随机效应和误差的分布(通常假定正态分布)。换句话说这些估计依赖于分布假定。此外，方差参数在 LMM 中是关键，例如前面提到若$\tilde{D}$已知，则固定效应β的估计可以用广义最小二乘方法来估计。因此本节研究方差参数的估计问题。由于只涉及到模型的二阶矩，因此可以用观测的某些二次函数来估计而不涉及分布问题。

Rao(1973)提出了一种现在称为最小范数二次无偏估计(Minimum Norm Quadratic Unbiased Estimator, MINQUE)方法来估计方差参数。为简述其方法，先看标准线性模型：$Y_{n\times 1}=X_{n\times p}\beta_{p\times 1}+e_{n\times 1}$，$Ee=0$，$Cov(e)=\sigma^2 I_n$。众所周知，方差参数$\sigma^2$的无偏估计为

$$\hat{\sigma}^2=\frac{Y'\left[I_n-X(X'X)^{-}X'\right]Y}{n-r},$$

其中$r=rank(X)$。

假设用Y的某二次函数作为σ^2的估计

$$\hat{\sigma}^2=Y'AY,$$

这里$A_{n\times n}$为某对称矩阵。由于要作为σ^2的估计，因此要求对所有Y，$Y'AY\geq 0$，因此要求$A\geq 0$。其次，我们希望此估计是无偏的。由于$E(Y'AY)=\beta'X'AX\beta+\sigma^2 tr(A)$，要满足无偏要求则必须有

$$X'AX=0 \text{且} tr(A)=1。$$

按照最小范数要求，在上面的约束下求A使得$tr(AA')=\min$。

采用 Lagrange 乘子法，定义 Lagrange 函数：

$$L(A,\Lambda_1,\lambda_2)=\frac{1}{2}tr(AA')+tr(X'AX\Lambda_1')+\lambda_2\left[1-tr(A)\right],$$

这里Λ_1为$p\times p$，λ_2为常数。

由以下公式

$$\frac{\partial tr(A)}{\partial A}=I,\ \frac{\partial tr(AA')}{\partial A}=2A,\ \frac{\partial tr(CAB)}{\partial A}=C'B',$$

有

$$\frac{\partial L(A,\Lambda_1,\lambda_2)}{\partial A}=A+X\Lambda_1X'-\lambda_2 I_n=0。$$

从而得$A=\lambda_2 I_n - X\Lambda_1 X'$，再由$X'AX=0$得$0=X'AX=\lambda_2 X'X - X'X\Lambda_1 X'X$，因此有$\Lambda_1=\lambda_2(X'X)^-$。代入得到$A=\lambda_2\left[I_n - X(X'X)^- X'\right]$。再由$tr(A)=1$，得到$\lambda_2=\dfrac{1}{n-rank(X)}$。故$A=\dfrac{I_n - X(X'X)^- X'}{n-rank(X)}$。可见，常见的$\hat{\sigma}^2$即为 MINQUE。

若误差为正态分布，则$Var(Y'AY)=2\sigma^4 tr(AA')$，因此最小化$tr(AA')$也意味着最小化估计$\hat{\sigma}^2$的方差。此时$Var(\hat{\sigma}^2)=\dfrac{2\sigma^4}{n-r}$。如果误差非正态分布，则$Var(Y'AY)$依赖于一些三阶和四阶矩。但直观上若$A$的元素增加$\rho$倍，其方差$Var(Y'AY)$也增加$\rho^2$倍，因此要使得$tr(AA')=\min$也是有道理的。

将上述 MINQUE 方法用于 LMM。考虑 LMM 的边际模型(见 7.1 节)

$$Y=X\beta+\eta,\quad E\eta=0,$$

这里$\eta=Zb+e$，$Z=\begin{pmatrix} Z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_K \end{pmatrix}$为$n\times(mK)$阶矩阵，

$$Cov(\eta)=\begin{pmatrix} Z_1 DZ_1' & 0 & \cdots & 0 \\ 0 & Z_2 DZ_2' & & \vdots \\ \vdots & \vdots & & 0 \\ 0 & 0 & \cdots & Z_K DZ_K' \end{pmatrix}+\sigma^2 I_n$$

$$=Z(I_K\otimes D)Z'+\sigma^2 I_n$$

先考察σ^2的 MINQUE 问题。考察二次型$Y'AY$，$A\geq 0$。此时

$$E(Y'AY)=\beta' X'AX\beta$$

$$+tr\left\{A\left[Z(I_K\otimes D)Z'+\sigma^2 I_n\right]\right\}。$$

要使其为σ^2的无偏估计要求满足

$X'AX=0$，$Z'AZ=0$，$tr(A)=1$。

定义$n\times(p+mK)$阶矩阵W如下

$$W=\begin{pmatrix}X & Z\end{pmatrix}=\begin{pmatrix}X_1 & Z_1 & 0 & 0\\ \vdots & 0 & \ddots & 0\\ X_K & 0 & 0 & Z_k\end{pmatrix}。$$

由于假设$A\geq 0$，故

$X'AX=0$，$Z'AZ=0\Leftrightarrow W'AW=0$。

类似上面的推导，此时σ^2的 MINQUE 为

$$\hat{\sigma}^2_{MINQUE}=\frac{Y'\left[I_n-W(W'W)^{-}W'\right]Y}{n-rank(W)}。$$

注：分子

$$Y'\left[I_n-W(W'W)^{-}W'\right]Y=\min_{v\in R^{p+mK}}\left\|Y-Wv\right\|^2$$

$$=\min_{\substack{\beta\in R^p\\ \gamma_1,\gamma_2\cdots\gamma_K\in R^m}}\sum_{i=1}^{K}\left\|Y_i-X_i\beta-Z_i\gamma_i\right\|^2,$$

在正态分布假定下有

$$Var(\hat{\sigma}^2_{MINQUE})=\frac{2\sigma^4}{n-rank(W)}。$$

现在来考察D的 MINQUE 问题。由于D为$m\times m$阶矩阵，因此将D按列拉成列向量，记作$vec(D)$，$vec(D)$为$m^2\times 1$向量。构造Y的二次函数$A(Y\otimes Y)$，这里A为$m^2\times n^2$矩阵。

注：相当于把D的所有m^2元素看成未知的，实际上D的未知只有$\frac{m(m+1)}{2}$。此外并没有考虑$D\geq 0$，所以对矩阵A没有加限制。因此所得 MINQUE 不一定能保证其非负定。

利用$\otimes$对加法有分配率与结合律，有

$$E\left[A(Y\otimes Y)\right]=A(X\beta\otimes X\beta)+AE(Zb\otimes Zb)$$
$$+AE(e\otimes e)。$$

再利用$\otimes$的性质

$$X\beta\otimes X\beta=(X\otimes X)(\beta\otimes\beta),$$
$$E(e\otimes e)=Evec(ee')=\sigma^2 vec(I_n),$$
$$E(Zb\otimes Zb)=E(Z\otimes Z)(b\otimes b)$$
$$=(Z\otimes Z)Evec(bb')$$
$$=(Z\otimes Z)vec(I_n\otimes D)$$

注意到$vec(I_n \otimes D)$为$vec(D)$的已知线性函数，即

$$vec(I_n \otimes D) = J \cdot vec(D),$$

这里J为$(nm)^2 \times m^2$的已知矩阵。因此要使得$E\left[A(Y \otimes Y)\right]$为$vec(D)$的无偏估计，则$A$满足

$$A(X \otimes X) = 0,\quad Avec(I_n) = 0,$$

$$A(Z \otimes Z)J = I_{m^2}。$$

因此在以上约束下，求A使得$tr(AA') = \min$。

分别定义$n^2 \times (p^2+1)$阶矩阵F及$(nm)^2 \times m^2$阶矩阵H如下：

$$F = \left[vec(I_n) \quad X \otimes X\right],$$

$$H = (Z \otimes Z)J$$

则上优化问题变为，求A使得$tr(AA') = \min$，满足以下约束

$$AF = 0,$$

$$AH = I_{m^2}。$$

构造 Lagrange 函数：

$$L(A, \Lambda_1, \Lambda_2) = \frac{1}{2}tr(AA') + tr(AF\Lambda_1') + tr\left[(I_{m_2} - AH)\Lambda_2'\right]。$$

则$\frac{\partial L}{\partial A} = A + \Lambda_1 F' - \Lambda_2 H' = 0$，从而得到$A = \Lambda_2 H' - \Lambda_1 F'$。由$0 = AF = \Lambda_2 H'F - \Lambda_1 F'F$，得$\Lambda_1 = \Lambda_2 H'F(F'F)^{-1}$，代入得$A = \Lambda_2 H'\left[I_{n^2m^2} - F(F'F)^{-1}F'\right]$。再由$AH = I_{m^2}$，有$\Lambda_2 = \left(H'\left[I_{n^2m^2} - F(F'F)^{-1}F'\right]H\right)^{-1}$。从而

$$A = \left(H'\left[I_{n^2m^2} - F(F'F)^{-1}F'\right]H\right)^{-1} H'\left[I_{n^2m^2} - F(F'F)^{-1}F'\right]。$$

注意到

$$F'F = \begin{pmatrix} n & vec'(X'X) \\ vec(X'X) & X'X \otimes X'X \end{pmatrix},$$

$$(F'F)^{-1} = \frac{1}{n-p}\begin{pmatrix} 1 & -vec'\left[(X'X)^{-1}\right] \\ -vec\left[(X'X)^{-1}\right] & (n-p)(X'X)^{-1} \otimes (X'X)^{-1} + vec\left[(X'X)^{-1}\right]vec'\left[(X'X)^{-1}\right] \end{pmatrix},$$

经过一系列繁琐推导$\hat{D}_{MINQUE}$有显示表达，且对于7.4，7.5小节的特殊情形有更简单表示。可以证明对7.4节的平衡数据其$\hat{D}_{MINQUE} = \hat{D}_{RML}$。

7.7 矩估计(Method of Moments)

上节介绍了估计方差参数D的 MINQUE，此方法并不需要对模型分布作何假定。本节来研究D的矩估计方法，同样也不需要对模型的分布作何种假定。矩估计方法(MM)的好处是，通常这样得到的估计都是无偏的且是一致的(consistent)。这里我们不妨假定σ^2的无偏估计已经给出，例如可以用上节的$\hat{\sigma}^2_{MINQUE}$。此方法最早由 Henderson(1953)提出，用于一种特殊且常见的线性混合效应模型。

称为方差分量模型(Variance components models，后面将提到)。Henderson 对此模型提出了三种方法I,II,III。由于 Henderson 方法III包括前两种。所以，本节的方法可以看成 Henderson 方法III的推广，见 Searle et.al(1992)。

矩估计方法一般有这样三个步骤。第一步，先用传统的最小二乘得到β的估计，即

$$\hat{\beta}_{OLS}=\left(\sum_{i=1}^{K}X_i'X_i\right)^{-1}\sum_{i=1}^{K}X_i'Y_i,$$

这样得到残差向量$\hat{e}_i=Y_i-X_i\hat{\beta}_{OLS},1\le i\le K$。然后利用残差向量$\hat{e}_i$来回归随机效应部分的$Z_i$，得到随机效应$b_i$的某种估计：

$$\hat{b}_i=\left(Z_i'Z_i\right)^{-}Z_i'\hat{e}_i,\ 1\le i\le K。$$

此处可能涉及到$Z_i'Z_i$不可逆，用广义逆代替。

第二步，计算一些交叉乘积和的期望值，例如$\sum_{i=1}^{K}\hat{b}_i\hat{b}_i'$的期望。

注意到$Y_i=X_i\beta+\eta_i$，这里$\eta_i=Z_ib_i+e_i$，且$E\eta_i=0,Cov(\eta_i)=\sigma^2I_{n_i}+Z_iDZ_i'$，从而有$\hat{e}_i=Y_i-X_i\hat{\beta}_{OLS}=\eta_i-\left(\sum_{j=1}^{K}X_j'X_j\right)^{-1}\sum_{j=1}^{K}X_j'\eta_j$。为简单起见，记$V_i=\sigma^2I_{n_i}+Z_iDZ_i'$，$N=\left(\sum_{j=1}^{K}X_j'X_j\right)^{-1}$，则$\hat{e}_i=\eta_i-N\sum_{j=1}^{K}X_j'\eta_j$。

$$E(\hat{e}_i\hat{e}_i') = E\left[\eta_i - N\sum_{j=1}^{K} X_j'\eta_j\right]\left[\eta_i - N\sum_{t=1}^{K} X_t'\eta_t\right]'$$

$$= V_i - V_i X_i NX_i' - X_i NX_i' V_i + X_i N\sum_{j=1}^{K} X_j' V_j X_j NX_i'$$

将$V_i = \sigma^2 I_{n_i} + Z_i DZ_i'$代入，合并整理得

$$E(\hat{e}_i\hat{e}_i') = \sigma^2(I_{n_i} - X_i NX_i') + Z_i DZ_i' - Z_i DZ_i' X_i NX_i' - X_i NX_i' Z_i DZ_i' + X_i N\sum_{j=1}^{K} X_j' Z_j DZ_j' X_j NX_i'$$

记$Z_i^+ = (Z_i'Z_i)^- Z_i'$，$J_i = Z_i^+ Z_i = (Z_i'Z_i)^- Z_i'Z_i$（为对称矩阵，如果满秩，则$J_i = I_m$），则得到

$$E\sum_{i=1}^{K}\left(\hat{b}_i\hat{b}_i'\right) = \sigma^2\sum_{i=1}^{K} Z_i^+(I_{n_i} - X_i NX_i')Z_i'^+ + \sum_{i=1}^{K} J_i DJ_i - \sum_{i=1}^{K} J_i DZ_i' X_i NX_i' Z_i'^+ - \sum_{i=1}^{n} Z_i^+ X_i NX_i' Z_i DJ_i + \sum_{i=1}^{K} Z_i^+ X_i N\sum_{j=1}^{K} X_j' Z_j DZ_j' X_j NX_i' Z_i'^+$$

第三步，采用矩估计方法(MM)

令

$$L = \sum_{i=1}^{K}\left(\hat{b}_i\hat{b}_i'\right) - \hat{\sigma}^2\sum_{i=1}^{K} Z_i^+(I_{n_i} - X_i NX_i')Z_i'^+,$$

为统计量，这里$\hat{\sigma}^2$为σ^2的无偏估计，则

$$EL = \sum_{i=1}^{K} J_i DJ_i - \sum_{i=1}^{K} J_i DZ_i' X_i NX_i' Z_i'^+ - \sum_{i=1}^{n} Z_i^+ X_i NX_i' Z_i DJ_i + \sum_{i=1}^{K} Z_i^+ X_i N\sum_{j=1}^{K} X_j' Z_j DZ_j' X_j NX_i' Z_i'^+$$

从而D的矩估计$\hat{D}_{MM}$，可以解如下矩阵方程：

$$L = \sum_{i=1}^{K} J_i\hat{D}_{MM}J_i - \sum_{i=1}^{K} J_i\hat{D}_{MM}Z_i'X_i NX_i'Z_i'^+ - \sum_{i=1}^{n} Z_i^+ X_i NX_i'Z_i\hat{D}_{MM}J_i + \sum_{i=1}^{K} Z_i^+ X_i N\sum_{j=1}^{K} X_j'Z_j\hat{D}_{MM}Z_j'X_j NX_i'Z_i'^+$$

记$R_{ij} = Z_i^+ X_i NX_j'Z_j$，则上方程变为

$$L = \sum_{i=1}^{K} J_i\hat{D}_{MM}J_i - \sum_{i=1}^{K} J_i\hat{D}_{MM}R_{ii}' - \sum_{i=1}^{n} R_{ii}\hat{D}_{MM}J_i + \sum_{i=1}^{K}\sum_{j=1}^{k} R_{ij}\hat{D}_{MM}R_{ij}'。$$

两边取vec，从而得到$vec(\hat{D}_{MM}) = F^{-1}vec(L)$，这里

$$F = \sum_{i=1}^{K} J_i \otimes J_i - \sum_{i=1}^{K} J_i \otimes R_{ii} - \sum_{i=1}^{n} R_{ii} \otimes J_i + \sum_{i=1}^{K}\sum_{j=1}^{k} R_{ij} \otimes R_{ij}。$$

如果对所以i，$Z_i'Z_i \geq aI_m$，则

$$tr(\sum_{i=1}^{K} Z_i^+ X_i N X_i' Z_i'^+) = tr\sum_{i=1}^{K} N X_i' Z_i'^+ Z_i^+ X_i$$
$$= tr\sum_{i=1}^{K} N X_i' (Z_i'Z_i)^{-1} X_i \leq \frac{p}{a},$$

可以证明$\frac{1}{K}\sum_{i=1}^{K} N X_i' Z_i'^+ Z_i^+ X_i = o_p(1)$。同理可以证明$\sum_{i=1}^{K} R_{ii}, \sum_{i=1}^{K}\sum_{j=1}^{K} R_{ij}$有界，从而

从而

$$K^{-1}F - I_m \otimes I_m = o_p(1)$$

$$K^{-1}L = K^{-1}\sum_{i=1}^{K}\left(\hat{b}_i\hat{b}_i'\right) - \hat{\sigma}^2 K^{-1}\sum_{i=1}^{K}(Z_i'Z_i)^{-1} + o_p(1),$$

故

$$\hat{D}_{MM} = K^{-1}\sum_{i=1}^{K}\left(\hat{b}_i\hat{b}_i'\right) - \hat{\sigma}^2 K^{-1}\sum_{i=1}^{K}(Z_i'Z_i)^{-1} + o_p(1)。$$

这也很好解释，因为当$K \to \infty$时，$\hat{\beta}_{OLS}$趋于真实的β，从而对$1 \leq i \leq K$

$$E\left(\hat{b}_i\hat{b}_i'\right) \approx E\left(Z_i^+\eta_i\eta_i'Z_i'^+\right) = \sigma^2(Z_i'Z_i)^{-1} + D。$$

7.8 随机效应的估计

先假定模型参数β, σ^2, D已知，来对第i个个体的随机效应b_i来进行估计。考虑条件期望$E\left(b_i \middle| Y_i\right)$，在正态分布假定下，由于

$$\begin{pmatrix} b_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ X_i\beta \end{pmatrix} \begin{pmatrix} D & DZ_i' \\ Z_iD & \sigma^2 I_{n_i} + Z_i D Z_i' \end{pmatrix}\right),$$

从而

$$E\left(b_i \middle| Y_i\right) = DZ_i'(\sigma^2 I_{n_i} + Z_i D Z_i')^{-1}(Y_i - X_i\beta)$$

因此，假定方差参数$\tilde{D} = \frac{D}{\sigma^2}$已知，则$\beta$可用广义最小二乘$\hat{\beta}$来估计，从而有$b_i$的估计：

$$\hat{b}_i = \tilde{D}Z_i'(I_{n_i} + Z_i\tilde{D}Z_i')^{-1}(Y_i - X_i\hat{\beta})。$$

由于$Z_i'(I_{n_i} + Z_i\tilde{D}Z_i')^{-1} = (I_m + Z_i'Z_i\tilde{D})^{-1}Z_i'$，因此

$$\hat{b}_i = \tilde{D}(I_m + Z_i'Z_i\tilde{D})^{-1}Z_i'(Y_i - X_i\hat{\beta})。$$

如果$D = 0$，没有随机效应，估计$\hat{b}_i = 0$。上式也是最优线性无偏预测(Best Linear Unbiased Predictor, BLUP)，见 Henderson(1963)。

考虑混合效应模型

$$Y = X\beta + Zb + e,$$

这里

$Cov(e) = \sigma^2 I_n$，$Cov(b) = \sigma^2 (I_K \otimes \tilde{D}) \equiv \sigma^2 H$，

$Cov(Y) = \sigma^2 (I_n + ZHZ') \equiv \sigma^2 \Sigma$。

考察b的线性无偏预测$\hat{b} = CY$，由无偏性有$CX = 0$。此外希望$E(\hat{b} - b)'(\hat{b} - b) = \min$。由于

$$\hat{b} - b = CY - b = (CZ - I)b + Ce,$$

$$E(\hat{b} - b)'(\hat{b} - b) = trCov(\hat{b} - b)$$
$$= \sigma^2 tr\left[CC' + (CZ - I)H(CZ - I)'\right]。$$

即在约束$CX = 0$下，使得

$$tr\left[CC' + (CZ - I)H(CZ - I)'\right] = \min。$$

构造 Lagrange 函数

$$L(C, \Lambda) = \frac{1}{2} tr\left[CC' + (CZ - I)H(CZ - I)'\right] - tr(CX\Lambda')，$$

由 $\frac{\partial L}{\partial C} = C + (CZ - I)HZ' - \Lambda X' = 0$ ，得

$C = (HZ' + \Lambda X')(I + ZH'Z)^{-1} = (HZ' + \Lambda X')\Sigma^{-1}$。

再由$CX = 0$得

$$\Lambda = -HZ'\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1},$$

因此，

$$C = \left[HZ' - HZ'\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\right]\Sigma^{-1}$$
$$= HZ'\Sigma^{-1}\left[I - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\right]。$$

从而

$$\hat{b} = HZ'\Sigma^{-1}\left[I - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\right]Y$$
$$= HZ'\Sigma^{-1}(Y - X\hat{\beta})$$

这里$\hat{\beta}$为(H，即Σ已知)广义最小二乘估计。

将$H = I_K \otimes \tilde{D}$代入，考察第$i$块有

$$\hat{b}_i = \tilde{D}Z_i'(I_{n_i} + Z_i\tilde{D}Z_i')^{-1}(Y_i - X_i\hat{\beta}),$$

此式与前面假定正态分布时推导一样。

此外，还可以采用下面的优化问题方式同时得到固定效应β及随机效应b_i的估计：

$$\min_{\beta, b_1, b_2, \cdots b_K} \sum_{i=1}^{K} \left(\left\| Y_i - X_i\beta - Z_i b_i \right\|^2 + b_i'\tilde{D}b_i \right)。$$

这是因为，给定β，当$b_i = (Z_i'Z_i + \tilde{D}^{-1})^{-1}Z_i'(Y_i - X_i\beta)$时达到最小。

再将b_i代入，得到关于β的优化问题

$$\min_{\beta}\sum_{i=1}^{K}(Y_i-X_i\beta)'G_i(Y_i-X_i\beta),$$

这里

$$G_i=\left[I_{n_i}-(Z_i'Z_i+\tilde{D}^{-1})^{-1}Z_i'\right]\left[I_{n_i}-(Z_i'Z_i+\tilde{D}^{-1})^{-1}Z_i'\right]。$$
$$+Z_i(Z_i'Z_i+\tilde{D}^{-1})^{-1}\tilde{D}^{-1}(Z_i'Z_i+\tilde{D}^{-1})^{-1}Z_i'$$

经化简有$G_i=(I_{n_i}+Z_i\tilde{D}Z_i')^{-1}$，这意味着当$\beta$为($\tilde{D}$已知)广义最小二乘估计时达到最小。

7.9 方差分量模型(Variance Component Models)

前面已经提到过，线性混合效应模型的一般形式可以写成

$$Y=X\beta+Zb+e,$$

Y为$n\times1$观测向量，$X_{n\times p}$为已知设计矩阵，$\beta_{p\times1}$未知为固定效应，$Z_{n\times q}$为已知设计矩阵，$b_{q\times1}$为随机效应，且设$Eb=0$，$Cov(b)=G$非负定，e为随机误差且与b独立，$Ee=0$，$Cov(e)=R$为正定矩阵。$Cov(Y)=ZGZ'+R$。，

前面几节，我们主要关注由 Laird & Ware(1982)发展起来的一类混合效应模型。这类混合效应模型针对每个个体建模，相当于$Z=diag(Z_1,Z_2,\cdots Z_K)$是对角块，其中$Z_i$为$n_i\times m$阶矩阵，$q=mK$，随机效应协方差阵$Cov(b)$也是对角块$diag(D_{m\times m},\underset{K个}{D_{m\times m}}\cdots D_{m\times m})$，即$G=I_K\otimes D_{m\times m}$。为简单起见假定$Cov(e)=R=\sigma^2I_n$，这里$n=\sum_{i=1}^{K}n_i$。

比这个更早的一类线性混合效应模型，其形式表现为随机效应部分Zb可以表示成r个互不相关(通常假定独立)的随机因子的叠加，即

$$Zb=\sum_{j=1}^{r}Z_jb_j,$$

这里每个Z_j为$n\times q_j$的已知矩阵，b_j为$q_j\times1$阶矩阵。记$q=\sum_{j=1}^{r}q_j$，通常假定$Eb_j=0$，$Cov(b_j)=\sigma_j^2I_{q_j}$。从而有：

$$Cov(Y)=\sum_{j=1}^{r}\sigma_j^2 Z_j Z_j'+R。$$

此模型将Y的波动(方差)分解为r个随机因子的波动和误差波动的叠加，因此称为方差分量模型(Variance Component Models)。相对于标准的线性混合效应模型来说，相当于

$$Z_{n\times q}=\begin{pmatrix}Z_1 & Z_2 & \cdots & Z_r\end{pmatrix},\ b_{q\times 1}=\begin{pmatrix}b_1\\ b_2\\ \vdots\\ b_r\end{pmatrix},$$

$$Cov(b)=diag\begin{pmatrix}\sigma_1^2 I_{q_1} & \sigma_2^2 I_{q_2} & \cdots & \sigma_r^2 I_{q_r}\end{pmatrix}。$$

为简单起见，以下不特别申明考虑$R=\sigma^2 I_n$。

为表示简单，记$\sigma_0^2=\sigma^2$，$Z_0=I_n$，则此时

$$\begin{aligned}Cov(Y)&=\sum_{j=1}^{r}\sigma_j^2 Z_j Z_j'+\sigma^2 I_n\\&=\sum_{j=0}^{r}\sigma_j^2 Z_j Z_j'\equiv V(\theta)\end{aligned},$$

这里$\theta=\left(\sigma_0^2,\sigma_1^2,\sigma_2^2,\cdots\sigma_r^2\right)'$。

例 7.9.1：单向分类随机效应模型

考虑模型$y_{ij}=\mu+\alpha_i+e_{ij}$，$j=1,2\cdots n_i$，$i=1,2,\cdots a$，这里$\mu$为固定效应，$\alpha_i$为随机效应。设$\alpha_i\ i.i.d\sim N(0,\sigma_\alpha^2)$，$e_{ij}\ i.i.d\sim N(0,\sigma^2)$。

记$n=\sum_{i=1}^{a}n_i$，$Y_{n\times 1}=(y_{11},\cdots y_{1n_1},\cdots y_{a1},\cdots y_{an_a})'$，$b_{a\times 1}=(\alpha_1,\alpha_2,\cdots\alpha_a)'$，$e_{n\times 1}=(e_{11},\cdots e_{1n_1},\cdots e_{a1},\cdots e_{an_a})'$，则写成标准形式(相当于$r=1$)

$$Y=E_n\mu+Zb+e,$$

这里，$Z_{n\times a}=\begin{pmatrix}E_{n_1} & & & \\ & E_{n_2} & & \\ & & \vdots & \\ & & & E_{n_a}\end{pmatrix}$。

$Cov(Y)=\sigma_a^2 diag(E_{n_1}E_{n_1}',\cdots E_{n_a}E_{n_a}')+\sigma^2 I_n$。

$V(\theta)=\theta_0 I_n+\theta_1 diag(E_{n_1}E_{n_1}',\cdots E_{n_a}E_{n_a}')$，其中$\theta_0=\sigma^2$，$\theta_1=\sigma_\alpha^2$且$X=E_n$。

对方差分量这类线性混合效应模型来说，前面关于参数估计的方法也都可以用于此模型。具体地，假定随机效应及误差为正态分布。此时对数似然函数为：

$$l=-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\log|V(\theta)| -\frac{1}{2}(Y-X\beta)'V^{-1}(\theta)(Y-X\beta)$$

。

$$l_\beta=\frac{\partial l}{\partial\beta}=X'V^{-1}(\theta)Y-X'V^{-1}(\theta)X\beta,$$

对$j=0,1\cdots r$，有

$$l_{\theta_j}=\frac{\partial l}{\partial\theta_j}=-\frac{1}{2}tr\left[V^{-1}(\theta)Z_jZ_j'\right] +\frac{1}{2}(Y-X\beta)'V^{-1}(\theta)Z_jZ_j'V^{-1}(\theta)(Y-X\beta)$$

令上面偏导数为零就得到似然方程组。

此外，还有另外一种等价的表示。

令

$$P(\theta)=V^{-1}(\theta)-V^{-1}(\theta)X\left[X'V^{-1}(\theta)X\right]^{-}X'V^{-1}(\theta),$$

由β的方程得

$$X'V^{-1}(\theta)X\beta=X'V^{-1}(\theta)Y,$$

也可写成

$$V^{-1}(\theta)(Y-X\beta)=P(\theta)Y。$$

关于θ的方程，对$j=0,1\cdots r$，有

$$tr\left[V^{-1}(\theta)Z_jZ_j'\right]=Y'P(\theta)Z_jZ_j'P(\theta)Y。$$

可见，后面关于方差参数的$r+1$个方程与β无关，可以解出θ，然后得到β的估计。

以上我们得到参数的极大似然(ML)估计。同前面的，也可以采用限制极大似然(RML)方法得到参数的 RMLE。设$rank(X_{n\times p})=r$，则存在$B_{n\times(n-r)}$，$rank(B)=n-r$，使得$B'X=0$。由$Y\sim N(X\beta,V(\theta))$，则$B'Y\sim N(0,B'V(\theta)B)$，这样由$BY$的似然函数(限制似然函数)，可以得到其似然方程，这只需要在前面的关于θ的似然方程里$Y\to B'Y$，$Z\to B'Z$，$X\to B'X=0$，$V(\theta)\to B'V(\theta)B$，$P(\theta)\to\left(B'V(\theta)B\right)^{-1}$代替即可。

对 $j=0,1\cdots r$，有

$$tr\left[(B'V(\theta)B)^{-1}B'Z_jZ_j'B\right]$$

$$=Y'B(B'V(\theta)B)^{-1}B'Z_jZ_j'B(B'V(\theta)B)^{-1}B'Y$$。

令$M(\theta)=B\left(B'V(\theta)B\right)^{-1}B'$，则上方程组写为

对 $j=0,1\cdots r$，

$$tr\left[M(\theta)Z_jZ_j'\right]=Y'M(\theta)Z_jZ_j'M(\theta)Y$$。

以下说明，$M(\theta)$并不依赖于B的选择，其恒等于前面的$P(\theta)$。即，限制极大似然方程为

$tr\left[P(\theta)Z_jZ_j'\right]=Y'P(\theta)Z_jZ_j'P(\theta)Y$，$j=0,1\cdots r$。

注：由于$\left[V^{-\frac{1}{2}}(\theta)X\right]'V^{\frac{1}{2}}(\theta)B=0$，因此

$rank\left(V^{-\frac{1}{2}}(\theta)X \quad V^{\frac{1}{2}}(\theta)B\right)=rank(X)+rank(B)=n$，

因此

$$I_n=P_{V^{-\frac{1}{2}}(\theta)X}+P_{V^{\frac{1}{2}}(\theta)B}$$

$$=V^{-\frac{1}{2}}(\theta)X\left[X'V^{-1}(\theta)X\right]^{-}X'V^{-\frac{1}{2}}(\theta)$$

$$+V^{\frac{1}{2}}(\theta)B\left[B'V^{-1}(\theta)B\right]^{-1}B'V^{\frac{1}{2}}(\theta)$$

即有$M(\theta)\equiv P(\theta)$。

例 7.9.2：平衡数据单向分类随机效应模型

考虑模型 $y_{ij}=\mu+\alpha_i+e_{ij}$，$j=1,2\cdots b$，$i=1,2,\cdots a$，这里$\mu$为固定效应，$\alpha_i$为随机效应。设$\alpha_i$ $i.i.d\sim N(0,\sigma_\alpha^2)$，$e_{ij}$ $i.i.d\sim N(0,\sigma^2)$。记 $n=ab$，$Y_{n\times1}=(y_{11},\cdots y_{1b},\cdots y_{a1},\cdots y_{ab})'$，$\alpha_{a\times1}=(\alpha_1,\alpha_2,\cdots\alpha_a)'$，$e_{n\times1}=(e_{11},\cdots e_{1b},\cdots e_{a1},\cdots e_{ab})'$，则写成标准形式(相当于$r=1$)

$$Y=X\mu+Z\alpha+e,$$

这里，$X_{n\times1}=E_a\otimes E_b=E_{ab}$，$Z_{n\times a}=I_a\otimes E_b$。

$Cov(Y)=\sigma_\alpha^2I_a\otimes E_bE_b'+\sigma^2I_{ab}$。

$V(\theta)=\theta_0I_{ab}+\theta_1I_a\otimes E_bE_b'$，其中$\theta_0=\sigma^2$，$\theta_1=\sigma_\alpha^2$。

$$V^{-1}(\theta)=\theta_0^{-1}I_{ab}-\frac{\theta_0^{-1}\theta_1}{\theta_0+\theta_1b}I_a\otimes E_bE_b',$$

$$V^{-1}(\theta)X=\frac{E_{ab}}{\theta_0+\theta_1b},\quad X'V^{-1}(\theta)X=\frac{ab}{\theta_0+\theta_1b}。$$

由μ的似然方程得

$$\mu=\overline{y}_{..}。$$

$$V^{-2}(\theta)=\theta_0^{-2}I_{ab}-\frac{2\theta_0^{-1}\theta_1+\theta_0^{-2}\theta_1^2 b}{(\theta_0+\theta_1 b)^2}I_a\otimes E_bE_b'$$

$$trV^{-1}(\theta)=ab\theta_0^{-1}(1-\frac{\theta_1}{\theta_0+\theta_1 b})$$

$$V^{-1}(\theta)\mathrm{Z}=\frac{1}{\theta_0+\theta_1 b}I_a\otimes E_b$$

$$V^{-1}(\theta)\mathrm{ZZ}'=\frac{1}{\theta_0+\theta_1 b}I_a\otimes E_bE_b'$$

$$V^{-1}(\theta)\mathrm{ZZ}'V^{-1}(\theta)=\frac{1}{(\theta_0+\theta_1 b)^2}I_a\otimes E_bE_b'$$

θ_0,θ_1的似然方程组为：

$$\begin{aligned}ab\theta_0^{-1}(1-\frac{\theta_1}{\theta_0+\theta_1 b})&=\theta_0^{-2}(Y-X\mu)'(Y-X\mu)\\ -\frac{2\theta_0^{-1}\theta_1+\theta_0^{-2}\theta_1^2 b}{(\theta_0+\theta_1 b)^2}&(Y-X\mu)'(I_a\otimes E_bE_b')(Y-X\mu)\end{aligned},$$

$$\frac{ab}{\theta_0+\theta_1 b}=(Y-X\mu)'\frac{(I_a\otimes E_bE_b')}{(\theta_0+\theta_1 b)^2}(Y-X\mu)。$$

由第二式，第一式化简为

$$\theta_0+\theta_1=\frac{(Y-X\mu)'(Y-X\mu)}{ab}。$$

将$\mu=\bar{y}_{..}$带入，得

$$\theta_0+\theta_1=\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2,$$

$$\theta_0+\theta_1 b=\frac{b}{a}\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2。$$

故得

$$\hat{\mu}=\bar{y}_{..},$$

$$\hat{\theta}_0=\frac{1}{a(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i.})^2,$$

$$\begin{aligned}\hat{\theta}_1&=\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2-\hat{\theta}_0\\&=\frac{1}{a}\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2-\frac{\hat{\theta}_0}{b}\end{aligned}$$

注：显然，似然方程组的解$\hat{\theta}_1$有可能取负值。因此，似然方程组的解不总是 ML 估计。当$\hat{\theta}_1$取负值时，方程组的解没有落到参数空间里。此时，似然函数的最大值在其边界$\theta_1=0$达到。即θ_1的 ML 估计为$\max\left\{\hat{\theta}_1,0\right\}$。

此外，对方差参数部分限制极大似然方程 $tr\left[P(\theta)Z_jZ_j'\right]=Y'P(\theta)Z_jZ_j'P(\theta)Y$， $j=0,1\cdots r$。还有另外一种表示。注意到由$P(\theta)$的定义有

$$P(\theta)V(\theta)P(\theta)=P(\theta)。$$

从而

$$\begin{aligned}tr\left[P(\theta)Z_jZ_j'\right]&=tr\left[P(\theta)V(\theta)P(\theta)Z_jZ_j'\right]\\&=tr\left[P(\theta)Z_jZ_j'P(\theta)V(\theta)\right]\\&=\sum_{i=0}^{r}tr\left[P(\theta)Z_jZ_j'P(\theta)Z_iZ_i'\right]\theta_i\end{aligned}。$$

这样，极大似然方程可以写为：

$$\left(tr\left[P(\theta)Z_iZ_i'P(\theta)Z_jZ_j'\right]\right)_{i,j=0}^{r}\theta=\begin{pmatrix}Y'P(\theta)Z_0Z_0'P(\theta)Y\\\vdots\\Y'P(\theta)Z_rZ_r'P(\theta)Y\end{pmatrix}。$$

注：上面这种形式可以在求解时构造θ的迭代求解——Anderson 迭代方法。

方差分量模型参数的 ML 估计及 RML 估计都要假定随机效应和误差的分布(通常假定正态分布)。换句话说这些估计依赖于分布假定。

与前面所说的混合效应模型一样，方差分量模型方差参数θ在模型中是关键。若θ已知，则固定效应β的估计可以用广义最小二乘方法来估计。接下来对于方差分量模型的方差参数介绍两种估计方法：ANOVA 估计方法及 MINQUE 方法。这些方法只涉及到模型的矩，而不涉及到分布问题。

设方差分量模型为：

$$Y=X\beta+\sum_{i=1}^{r}Z_ib_i+e。$$

为表示简单，记$\sigma_0^2=\sigma^2$，$Z_0=I_n$，$b_0=e$则此时模型为$Y=X\beta+\sum_{i=0}^{r}Z_ib_i$

$$Cov(Y)=\sum_{j=0}^{r}\sigma_j^2Z_jZ_j'\equiv V(\theta)，$$

这里$\theta=\left(\sigma_0^2,\sigma_1^2,\sigma_2^2,\cdots\sigma_r^2\right)'$。

假设有$r+1$个二次型$s_i = Y'A_iY$（假定$A_i \geq 0$），$i = 0,1,\cdots,r$。则

$$Es_i = EY'A_iY$$
$$= tr\left[A_iV(\theta)\right] + \beta' X'A_iX\beta$$

若A_i满足$X'A_iX = 0$。这样有

$$Es_i = EY'A_iY = tr\left[A_iV(\theta)\right]$$
$$= tr\left[A_i\sum_{j=0}^{r}\theta_j Z_j Z_j'\right] = \sum_{j=0}^{r} tr(Z_j'A_iZ_j)\theta_j$$

记$S = (s_0, s_1, \cdots s_r)'$，$C = \left(C_{ij} = tr(Z_j'A_iZ_j)\right)$，$0 \leq i, j \leq r$，则写成矩阵形式有：

$$ES = C\theta$$

从而由矩估计方法得到θ的估计$\hat{\theta}$

$$S = C\hat{\theta},$$

即

$$\hat{\theta} = C^{-1}S,$$

若C可逆。

这样，一个关键的问题是：满足$X'A_iX = 0$ $i = 0,1,\cdots,r$ 的这$r+1$个$\{A_i\}$是否存在？

Henderson 指出，可以选取

$$A_0 = I_n - P_{(X,Z_1,\cdots Z_r)},$$

对$i = 1,2,\cdots,r$

$$A_i = \left[I_n - P_{(X,Z_1,\cdots Z_{i-1})}\right] - \left[I_n - P_{(X,Z_1,\cdots Z_i)}\right]。$$
$$= P_{(X,Z_1,\cdots Z_i)} - P_{(X,Z_1,\cdots Z_{i-1})}$$

注：之所以称为 ANOVA 方法是因为也是利用 Y的一些平方和来估计。

上面的方法是刚好选取$r+1$个，C是方阵。事实上也可以超过$r+1$个，此时C不再是方阵，则θ的估计$\hat{\theta}$为最小二乘解：

$$\hat{\theta} = (C'C)^{-1}C'S。$$

再转到 MINQUE 方法。考察方差分量的已知线性函数$c'\theta$，考虑用某二次型$Y'AY$来估计$c'\theta$，要求A对称且满足$AX = 0$。此时

$$EY'AY = tr\left[AV(\theta)\right] = \sum_{j=0}^{r} tr(Z_j'AZ_j)\theta_j,$$

要满足无偏性，所以对 $j = 0,1,\cdots,r$

$$tr(Z_j'AZ_j)=c_j。$$

另一方面，若对$j=0,1,\cdots,r$，b_j都已知(为$q_j\times 1$向量)，则θ_j的估计应该用$\frac{b_j'b_j}{q_j}$，从而$c'\theta$的估计为$\sum_{j=0}^{r}c_j\frac{b_j'b_j}{q_j}=b'\Delta b$，这里$\Delta=diag\left(\frac{c_0}{q_0}I_{q_0},\cdots,\frac{c_r}{q_r}I_{q_r}\right)$，$b=(b_0',b_1',\cdots,b_r')'$。

记$Z=(Z_0,Z_1,\cdots,Z_r)$，则$Y=X\beta+Zb$，从而有$Y'AY=b'Z'AZb$（假设$AX=0$）。从而欲使得$Y'AY$是好的估计，对一切b有$b'Z'AZb$与$b'\Delta b$很接近。若用某种范数来度量，则选择A使得$\|Z'AZ-\Delta\|=\min$。注意到A满足以下条件：

$$AX=0$$

$$j=0,1,\cdots,r，\ tr(Z_j'AZ_j)=c_j。$$

对已知线性函数$c'\theta$，若估计$Y'AY$满足上约束条件且使得$\|Z'AZ-\Delta\|=\min$，则称$Y'AY$为$c'\theta$的最小范数二次无偏估计(MINQUE)。

C.R.Rao 选择加权欧式范数。设权矩阵(已知)

$$W=diag(w_0^2I_{q_0},w_1^2I_{q_1},\cdots w_r^2I_{q_r}),$$

令

$$F=W^{\frac{1}{2}}(Z'AZ-\Delta)W^{\frac{1}{2}}$$

则加权范数$\|Z'AZ-\Delta\|^2=tr(F'F)$。利用矩阵$A$满足的约束条件，可得

$$tr(F'F)=tr(AV(\theta_W))^2-tr(\Delta W)^2,$$

这里$V(\theta_W)=\sum_{j=0}^{r}w_j^2Z_jZ_j'>0$，$\theta_W=(w_0^2,w_1^2,w_2^2,\cdots w_r^2)'$。

从而，MINQUE 估计问题就归结为求下述极值问题：

$$\min tr(AV(\theta_W))^2$$

$$\begin{cases} AX=0, \\ tr(AZ_iZ_i')=c_i,i=0,1,\cdots,r \end{cases}。$$

定理 7.9.1：上极值问题的解为

$$A^*=\sum_{j=0}^{r}\lambda_iP(\theta_W)Z_iZ_i'P(\theta_W),$$

其中

$P(\theta_W)=V^{-1}(\theta_W)-V^{-1}(\theta_W)X\left[X'V^{-1}(\theta_W)X\right]^{-}X'V^{-1}(\theta_W)$，
$\lambda=(\lambda_0,\lambda_1,\cdots,\lambda_r)'$满足矩阵方程

$$\left(tr\left[P(\theta_W)Z_iZ_i'P(\theta_W)Z_jZ_j'\right]\right)_{i,j=0}^{r}\lambda=c。$$

这样，由上定理 $c'\theta$ 的 MINQUE 为 $c'\hat{\theta}=Y'A^*Y$。事实上这里$\hat{\theta}$为下方程组的解：

$$\left(tr\left[P(\theta_W)Z_iZ_i'P(\theta_W)Z_jZ_j'\right]\right)_{i,j=0}^{r}\theta=\begin{pmatrix}Y'P(\theta_W)Z_0Z_0'P(\theta_W)Y\\ \vdots\\ Y'P(\theta_W)Z_rZ_r'P(\theta_W)Y\end{pmatrix}$$

注：由$c'\hat{\theta}=Y'A^*Y$，有

$$c'\hat{\theta}=\lambda'\left(tr\left[P(\theta_W)Z_iZ_i'P(\theta_W)Z_jZ_j'\right]\right)_{i,j=0}^{r}\hat{\theta},$$

$$\begin{aligned}Y'A^*Y&=\sum_{j=0}^{r}\lambda_iY'P(\theta_W)Z_iZ_i'P(\theta_W)Y\\&=\lambda'\begin{pmatrix}Y'P(\theta_W)Z_0Z_0'P(\theta_W)Y\\ \vdots\\ Y'P(\theta_W)Z_rZ_r'P(\theta_W)Y\end{pmatrix},\end{aligned}$$

再由c的任意性，可得。