第六章　秩相关和稳健回归

这一章是定量变量间的相互依赖关系. 前4节是有关变量的相关关系,包括两个变量之间的秩相关分析和多变量之间的协同关系, 后3节是几种稳健回归和分位数回归.

§6.1　Spearman秩相关检验

基本理论

设量为n的样本$(X,Y) = \{(X_1,Y_1),\cdots,(X_n,Y_n)\} \overset{i.i.d.}{\sim} F(x,y)$. 假设检验问题为

$$H_0: X与Y不相关 \leftrightarrow H_1: X与Y正相关. \tag{6.1}$$

对上面的假设检验问题, 当H_1成立时, 说明随着X的增加Y也增加, 即X与Y具有某种同步性. 在参数推断中, 两个随机变量之间的相关性常通过相关系数度量, Pearson相关系数定义为

$$r(X,Y) = \frac{\sum_{i=1}^{n}[(X_i-\bar{X})(Y_i-\bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}},$$

其中, $-1 < r < +1$. 当$r > 0$时, 表示X与Y正相关; $r < 0$时, 表示X与Y负相关; $r = 0$时, 表示X与Y不相关.

在学生IQ和EQ数据中, 如果使用常规的Pearson相关系数, 会发现在观测到的学生中, IQ与EQ的相关性非常高, 达到0.9184, 这似乎是学生学业优异处世能力一定强的有力佐证. 如果做散点图, 可以清晰地观察到两组数据本质上是没有关系的, 导致两组数据呈现高度相关性的一个直接原因是出现了一名IQ和EQ都很高的特殊的学生, 这名学生的情况和大部分学生的特点不同, 放在一个分布之下进行分析是不合理的. 是否有其他的方法在我们肉眼观察不到的时候能够将这种异常的情况显现出来(比如数据量很大, 作图并不实用)? 剔除这些影响数据整体关系的干扰元素, 将主体相关性比较客观地计算出来, 这就是本节和6.2节介绍的秩相关系数.

令R_i表示X_i在$(X_1,X_2,\cdots,X_n)$中的秩, Q_i表示Y_i在$(Y_1,Y_2,\cdots,Y_n)$ 中的秩, 如果X_i 与Y_i 具有同步性,那么R_i与Q_i也表现出同步性, 反之亦然. 仿照样本相关系数$r(X,Y)$的计算方法, 定义秩之间的一致性, 因而有了Spearman相关系数：

$$r_S=\frac{\sum_{i=1}^{n}\left[\left(R_i-\frac{1}{n}\sum_{i=1}^{n}R_i\right)\left(Q_i-\frac{1}{n}\sum_{i=1}^{n}Q_i\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(R_i-\frac{1}{n}\sum_{i=1}^{n}R_i\right)^2}\sqrt{\sum_{i=1}^{n}\left(Q_i-\frac{1}{n}\sum_{i=1}^{n}Q_i\right)^2}}. \tag{6.2}$$

注意到

$$\sum_{i=1}^{n}R_i=\sum_{i=1}^{n}Q_i=\frac{n(n+1)}{2},$$
$$\sum_{i=1}^{n}R_i^2=\sum_{i=1}^{n}Q_i^2=\frac{n(n+1)(2n+1)}{6},$$

因此r_S可以简化为

$$r_S=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i-Q_i)^2. \tag{6.3}$$

参数统计中用t检验来进行相关性检验, 在零假设之下, 也可以类似的定义T检验统计量:

$$T=r_S\sqrt{\frac{n-2}{1-r_S^2}}. \tag{6.4}$$

该统计量在零假设之下服从$\nu=n-2$的t分布, 当$T>t_{\alpha,\nu}$时, 表示两变量有相关关系, 反之则无. 如果数据中有重复数据, 可以采用平均秩法定秩, 当结不多时, 仍然可以使用r_S定义秩相关系数, T检验仍然可以使用.

例6.1 有研究发现, 学生的中学英语学习成绩与大学英语学习成绩之间有相关关系,现收集某大学部分学生一年级英语期末成绩, 与其高考英语成绩进行比较,调查12位学生的结果如表6.1 所示, 用Spearman秩相关系数检验.

表6.1 学生高考英语成绩和大学英语成绩比较表

高考成绩x	65	79	67	66	89	85	84	73	88	80	86	75
大学成绩y	62	66	50	68	88	86	64	62	92	64	81	80

假设检验问题为

H_0:学生高考英语成绩与大学英语成绩不相关,

H_1:学生高考英语成绩与大学英语成绩相关.

将表6.1中学生的分数定秩后如表6.2所示.

表6.2 学生高考成绩和大学成绩秩计算表

x秩	1	6	3	2	12	9	8	4	11	7	10	5
y秩	2.5	6	1	7	11	10	4.5	2.5	12	4.5	9	8
$R_i - Q_i$	-1.5	0	2	-5	1	-1	3.5	1.5	-1	2.5	1	-3

计算秩差的平方和为

$$\sum (R_i - Q_i)^2 = (-1.5)^2 + \cdots + (-3)^2 = 65.$$

由式(6.3)得

$$r_S = 1 - \frac{6 \times 65}{12^3 - 12} = 1 - 0.2273 = 0.7727.$$

由式(6.4)得

$$T = 0.7727\sqrt{\frac{12-2}{1-0.7727^2}} = 3.8494.$$

实测$T = 3.8494 > t_{0.01,10} = 3.169$, 接受$H_1$假设, 认为学生英语高考成绩与大学成绩有关. R检验程序如下：

```
score.highschool=c(65,79,67,66,89,85,84,73,88,80,86,75)
score.univ=c(62,66,50,68,88,86,64,62,92,64,81,80)

cor.test(score.highschool, score.univ, meth="spearman")
        Spearman's rank correlation rho
data:  score.highschool and score.univ
S = 65.2267, p-value = 0.003265
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7719346
```

程序中的rho就是spearman相关系数. $S = 65.2267$表示秩平方差$\sum (R_i - Q_i)^2$.

关于r_S在零假设的分布有下面定理.

定理6.1 在零假设之下, Spearman秩相关系数分布满足：

(1) $E_{H_0}(r_S) = 0, \mathrm{var}_{H_0}(r_S) = \dfrac{1}{n-1}$;

(2) 关于原点0对称.

证明 在零假设之下, $(R_1, R_2, \cdots, R_n)$在空间$R = \{(i_1, i_2, \cdots, i_n) : (i_1, i_2, \cdots, i_n)$是$(1, 2, \cdots, n)$的排列$\}$上服从均匀分布. 注意到$r_S$的分布只与$\sum_{i=1}^{n}(R_i - Q_i)^2$有关, 因此, 首先计算

$$\sum_{i}^{n}(R_i - Q_i)^2 = \frac{n(n+1)(2n+1)}{3} - 2\sum_{i=1}^{n}(iR_i).$$

由推论(2.3)易知

$$E_{H_0}\left(\sum_{i}^{n}(R_i - Q_i)^2\right) = \frac{n(n^2-1)}{6}, \quad \operatorname{var}_{H_0}\left(\sum_{i}^{n}(R_i - Q_i)^2\right) = \frac{n^2(n+1)^2(n-1)}{36}.$$

下面证明对称性.

在H_0下, $(R_1, R_2, \cdots, R_n)$与$(n+1-R_1, \cdots, n+1-R_i)$同分布, 即$(R_1, R_2, \cdots, R_n) \overset{\mathrm{d}}{=} (n+1-R_1, \cdots, n+1-R_i)$. 于是在$H_0$下,

$$\begin{aligned}\sum_{i=1}^{n}(iR_i) - \frac{n(n+1)^2}{4} &= \sum_{i=1}^{n} i\left[\frac{n+1}{2} - (n+1-R_i)\right] \\ &= \sum_{i=1}^{n} i\left(\frac{n+1}{2} - R_i\right) \\ &= \frac{n(n+1)^2}{4} - \sum_{i=1}^{n}(iR_i).\end{aligned}$$

即统计量$\sum_{i}^{n}(R_i - Q_i)^2$在$H_0$下关于

$$E_{H_0}\left(\sum_{i}^{n}(R_i - Q_i)^2\right) = \frac{n(n+1)(2n+1)}{3} - 2\frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{6}$$

对称.

根据定理6.1可以方便地构造Spearman秩相关系数零分布表. 如果令$\alpha(2)$表示双边假设H_0 : X与Y不相关$\leftrightarrow$ H_1 : X与Y相关的显著性水平, $\alpha(1)$则为单边假设H_0 : X与Y不相关$\leftrightarrow$ H_1 : X与Y正相关的显著性水平. 经上面分析, 当$r_S \geqslant c_{\alpha(1)}$(双边时为$r_S \geqslant c_{\alpha(2)}$或者$r_S \leqslant c_{\alpha(2)}$)时拒绝$H_0$.

当n较大时, 霍特林(H·Hotelling)等人于1936年证明, Spearman秩相关系数有如下的大样本性质:

当$n \to \infty$时,

$$\sqrt{n-1}\, r_S \xrightarrow{\mathcal{L}} N(0,1).$$

因此在大样本时, 可用正态近似.

当X或Y样本中有结存在时, 可按平均秩法定秩, 相应的Spearman相关系数

$$r^*=\frac{\dfrac{n(n^2-1)}{6}-\dfrac{1}{12}\left[\sum_{i=1}^{n}(\tau_i^3(x)-\tau_i(x))+\sum_{i=1}^{n}(\tau_i^3(y)-\tau_i(y))\right]-\sum_{i=1}^{n}(R_i-Q_i)^2}{2\sqrt{\left[\dfrac{n(n^2-1)}{12}-\dfrac{1}{12}\sum_{i=1}^{n}(\tau_i^3(x)-\tau_i(x))\right]\left[\dfrac{n(n^2-1)}{12}-\dfrac{1}{12}\sum_{i=1}^{n}(\tau_i^3(y)-\tau_i(y))\right]}}$$

作为检验统计量, 其中$\tau_i(x)$,$\tau_i(y)$分别表示X,Y样本中的结统计量.

当结的长度较小时, 关于r^*的零分布仍可用无结时的零分布近似, 当n较大时, 也可用下面的极限分布:

$$r^*\sqrt{n-1}\xrightarrow{\mathcal{L}}N(0,1)$$

进行大样本检验.

关于Spearman秩相关系数对传统的样本相关系数的效率比较, 霍特林(H·Hotelling)和帕勃斯特(M·R·Pabst) 于1936年估算Spearman 的等级相关系数的效率约为Pearson相关系数的91%; 关于后一种相关系数检验,巴塔查里亚(Bhattacharyya)等在1970年指出: 当分布函数$F(x,y)$为$N(\mu_1,\mu_2,\sigma_1,\sigma_2;\rho)$时, Spearman 秩相关系数对样本相关系数$r(X,Y)$的渐近相对效率为$\dfrac{9}{\pi^2}\approx 0.912$. 这些结果说明在正态分布假定之下, 二者在效率方面是等价的. 但它们的效率都比较低, 而对于非正态分布的数据, 采用秩相关比较合适.

例6.2(IQ和EQ数据)　计算Spearman相关系数为: $r^*=0.3032$, 检验p值为0.097, 所以不能拒绝零假设, 不支持学生IQ与EQ强相关性存在.

例6.3(例6.1续)　因为数据中有秩, 因而按照有结情况计算:

$$\sum(R_i-Q_i)^2=(-1.5)^2+\cdots+(-3)^2=65.$$

相应的Spearman相关系数为

$$r^*=0.7719,$$

$$r^*\sqrt{n-1}=2.56.$$

标准正态分布$\alpha=0.05$, 对应的分位数$c_\alpha=1.96<2.56$, 所以, 拒绝零假设,接受H_1假设, 认为学生英语高考成绩与大学成绩有关,两种检验结果一致.

§6.2 Kendall τ相关检验

考虑假设检验问题:

$$H_0: X与Y不相关\leftrightarrow H_1: X与Y正相关.$$

肯德尔(Kendall)于1938年提出另一种与Spearman秩相关相似的检验法. 他从两变量(x_i,y_i), $(i=1,2,\cdots,n)$ 是否协同一致的角度出发检验两变量之间是否存在相关性. 首先引入协同的概念, 假设有n对观测值(x_1,y_1), (x_2,y_2), $\cdots$, (x_n,y_n),如果

乘积$(x_j-x_i)(y_j-y_i)>0,\ \forall j>i, i,j=1,2,\cdots,n$, 称数对$(x_i,y_i)$ 与(x_j,y_j) 满足协同性(concordant).或者说,它们的变化方向一致. 反之, 如果乘积$(x_j-x_i)(y_j-y_i)<0,\ \forall j>i, i,j=1,2,\cdots,n$, 则称该数对不协同(disconcordant), 表示变化方向相反. 也就是说, 协同性测量了前后两个数对的秩大小变化同向还是反向, 若前一对均比后一对秩小, 则说明前后数对具有同向性; 反之, 若前一对的秩比后一对大, 则前后两对数对(x_i,y_i)与(x_j,y_j)反向.

全部数据所有可能前后数对共有$\binom{n}{2}=n(n-1)/2$对. 如果用N_{c}表示同向数对的数目, N_{d}表示反向数对的数目, 则$N_{\mathrm{c}}+N_{\mathrm{d}}=n(n-1)/2$, Kendall相关系数统计量由二者的平均差定义, 如下所示:

$$\tau=\frac{N_{\mathrm{c}}-N_{\mathrm{d}}}{n(n-1)/2}=\frac{2S}{n(n-1)}. \tag{6.5}$$

式中, $S=N_{\mathrm{c}}-N_{\mathrm{d}}$, 若所有数对协同一致, 则$N_{\mathrm{c}}=n(n-1)/2$, $N_{\mathrm{d}}=0,\ \ \tau=1$, 表示两组数据正相关; 若所有数对全反向, 则$N_{\mathrm{c}}=0,\ \ N_{\mathrm{d}}=n(n-1)/2,\ \ \tau=-1$, 表示两组数据负相关; Kendall τ为零, 表示数据中同向和反向的数对势力均衡, 没有明显的趋势, 这与相关性的含义是一致的. 总之, Kendall τ在$-1\leqslant\tau\leqslant+1$之间, 反映了两组数据的变化一致性. 该统计量是肯德尔(Kendall)于1938年提出的, 因而称为Kendall τ检验统计量. H_0的拒绝域为τ取大值. 卡塞马克(Kaarsemaker)和温加尔登(Wijingaarden)于1953年给出了Kendall τ检验的零分布.

另外, 我们注意到, 如果定义

$$\mathrm{sign}((X_1-X_2)(Y_1-Y_2))=\begin{cases}1, & (X_1-X_2)(Y_1-Y_2)>0,\\ 0, & (X_1-X_2)(Y_1-Y_2)=0,\\ -1, & (X_1-X_2)(Y_1-Y_2)<0;\end{cases}$$

则

$$\tau=\frac{2}{n(n-1)}\sum_{1\leqslant i<j\leqslant n}\mathrm{sign}((x_i-x_j)(y_i-y_j)).$$

式中, $\mathrm{sign}((x_1-x_2)(y_1-y_2))$是$P((x_1-x_2)(y_1-y_2)>0)$的核估计量, 因而$\tau$是$U$统计量. 用$U$统计量的方法, 不难证明下面的定理.

定理6.2 在零假设H_0成立时,

(1) $E_{H_0}(\tau)=0,\qquad \mathrm{var}_{H_0}(\tau)=\dfrac{2(2n+5)}{9n(n-1)}$;

(2) 关于原点0对称.

当H_1成立时, $E\tau>0$. 于是, 当样本量n很大时, 根据U统计量的性质, 在H_0下

可以证明, 当$n \to \infty$时, 有

$$\tau\sqrt{\frac{9n(n-1)}{2(2n+5)}} \xrightarrow{\mathcal{L}} N(0,1).$$

实际中, 不失一般性, 假定x_i已从小到大或从大到小排序, 因此协同性问题就转化为y_i 秩的变化. 令$d_1, d_2, \cdots, d_n$为$y_1, y_2, \cdots, y_n$的秩, 因而x, y的秩形成$(1, d_1), (2, d_2), \cdots, (n, d_n)$; $\forall 1 \leqslant i \leqslant n$,记

$$p_i = \sum_{j>i} I(d_j > d_i), i = 1, 2, \cdots, n; \quad q_i = \sum_{j>i} I(d_j < d_i), i = 1, 2, \cdots, n.$$

令$P = \sum_{i=1}^{n} p_i, Q = \sum_{i=1}^{n} q_i$; 则Kendall τ统计量的值为$K = \dfrac{P-Q}{n(n-1)/2}$. 也就是说, 对每一个$y_i$求当前位置后比$y_i$大的数据的个数, 将这些数相加所得就是$N_{\mathrm{c}}$. 同理可以计算$N_{\mathrm{d}}$. 具体计算如例6.4.

例6.4 现在想研究体重和肺活量的关系, 调查某地10名女初中生的体重和肺活量的数据如表6.3所示, 进行相关性检验.

表6.3 学生体重和肺活量比较表

学生编号	1	2	3	4	5	6	7	8	9	10
体重(x)	75	95	85	70	76	68	60	66	80	88
肺活量(y)	2.62	2.91	2.94	2.11	2.17	1.98	2.04	2.20	2.65	2.69
肺活量秩	6	9	10	3	4	1	2	5	7	8

解 假设检验问题为

H_0: 体重和肺活量没有相关关系,

H_1: 体重和肺活量有相关关系.

计算每个变量的秩如下表：

表6.4 体重从小到大排序和肺活量对应的秩数据表

学生代号	7	8	6	4	1	5	9	3	10	2
体重(x)顺序	1	2	3	4	5	6	7	8	9	10
肺活量(y)对应秩	2	5	1	3	6	4	7	10	8	9

N_{c}与N_{d}的求解方法如下：

$$N_{\mathrm{c}} = 38, \quad N_{\mathrm{d}} = 7, \quad S = N_{\mathrm{c}} - N_{\mathrm{d}} = 31;$$
$$n = 10, \quad n(n-1) = 10(10-1) = 90.$$

表6.5 Kendallτ数对求秩表

秩 (x_i, y_i)	N_{c}	N_{d}
1 2	8	1
2 5	5	3
3 1	7	0
4 3	6	0
5 6	4	1
6 4	4	0
7 7	3	0
8 10	0	2
9 8	1	0
10 9	0	0
	38	7

由式(6.5)得

$$\tau = \frac{2 \times 31}{90} = 0.6889.$$

R程序如下：

```
cor.test(Weight,Lung,meth="kendall")
        Kendall's rank correlation tau
data:  Weight and Lung
T = 38, p-value = 0.004687
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.6888889
```

p值很小, 接受H_1, 认为体重与肺活量有关, 体重重的学生,肺活量也大.

若x_i或y_i有相等秩时, 用平均秩计算各自的秩, Kendall的τ公式校正如下：

$$\tau = \frac{S}{\sqrt{n(n-1)/2 - T_x}\sqrt{n(n-1)/2 - T_y}}.$$

式中, $T_x = \dfrac{1}{2}\sum^{g_x}(\tau_x^3 - \tau_x)$, $T_y = \dfrac{1}{2}\sum^{g_y}(\tau_y^3 - \tau_y)$, τ_x, τ_y 分别为$\{x_i\}, \{y_i\}$的结长, g_x, g_y分别为两变量中结的个数.

关于Kendall τ的效率,Bhattacharyya等人于1970年指出, 两者间的ARE为$\dfrac{9}{\pi^2} \approx 0.912$. 有人也将Spearman相关系数和$\tau$做了比较, 就皮特曼(Pitman)的ARE而言, 对所有的总体分布ARE$(r_S, \tau) = 1$.这也表明两者对于样本相关系数的ARE是相同的.

莱曼(Lehmann) (1975) 发现, 对于所有的总体分布有$0.746 \leqslant \mathrm{ARE}(r_S, r) < \infty$. 而对于一种形式的备择假设, 科尼金(Konijn) (1956)给出了表6.6所示结果.

表6.6 相关系数r_S的效率

总体分布	正态	均匀	抛物	重指数
$\mathrm{ARE}(r_S, \tau)$	0.912	1	0.857	1.266

§6.3 多变量Kendall协和系数检验

前两节所介绍的Spearman和Kendall τ两种检验方法都是针对两变量的相关性,这种相关的概念可以延拓至多变量间的相关. 比如,在实际问题中人们感兴趣的是几个变量之间是否具有同步或相关性,如为了诊断病情, 通常病人要做许多项检查, 这些结果彼此之间是否存在着相关? 歌手大奖赛上, 有诸多评委对歌手进行打分, 就同一个歌手而言, 不同评委之间意见是否是一致的呢? 也就是说, 从平均的意义来看,某个歌手被某个专家给予高分, 是否意味着其他专家也对他打了高分呢? 肯德尔(Kendall)和巴宾顿(Babington) 于1939年提出的多变量协和系数检验(concordance of variables), 就是针对这类问题的. 变量间的协和系数检验是以多变量秩检验为基础所建立起来的.

假设有k个变量$\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_k$, 每个变量有$n$个观测值, 设第$j$个变量$\boldsymbol{X}_j = (X_{1j}, X_{2j}, \cdots, X_{nj})$, 假设检验问题为

$$H_0 : k\text{个变量不相关} \leftrightarrow H_1 : k\text{个变量相关}. \tag{6.6}$$

记R_{ij}为X_{ij}在$(X_{1_j}, X_{2_j}, \cdots, X_{n_j})$的秩, 表示成如下数据表形式:

表6.7 多变量的秩表示

	变量1	变量2	$\cdots$	变量k	和
秩	R_{11}	R_{12}	$\cdots$	R_{1k}	$R_{1\cdot}$
	R_{21}	R_{22}	$\cdots$	R_{2k}	$R_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	R_{n1}	R_{n2}	$\cdots$	R_{nk}	$R_{n\cdot}$

在H_0成立下, 各个变量应没有相关性, 因而从每一行来看, 各行秩和理应相差不大; 但在H_1下, 由于各变量有一致性, 因而存在某一行的秩和较大, 也存在某一

行的秩和很小. 在H_1下, 各行向量的秩和可能相差很大, 如果记$R_{i\cdot}=\sum_{j=1}^{k}R_{ij},(i=1,2,\cdots,k)$, 所有秩和$R_{\cdot\cdot}=\sum_{i=1}^{n}\sum_{j=1}^{k}R_{ij}=kn(n+1)/2$, 则可用统计量

$$\mathrm{S}=\sum_{i=1}^{n}\left(R_{i\cdot}-\frac{1}{n}\sum_{i=1}^{n}R_{i\cdot}\right)^2$$

检验假设。另一方面，如果各个变量每一个排名秩完全一致，那么每个变量(j)上对每个对象i 的秩都是相同的，秩和是$1k,2k,3k,\cdots,nk$的某种排列。在这种完全一致的情况下，每个秩和与平均值$k(n+1)/2$的偏差平方和为

$$\mathrm{T}=\sum_{i=1}^{n}\left(ik-k(n+1)/2\right)^2$$

在零假设之下, 我们有

$$\begin{aligned}
\mathrm{SST}&=\frac{1}{k}\mathrm{T}=\sum\sum R_{ij}^2-R_{\cdot\cdot}^2/nk\\
&=k(1^2+2^2+\cdots+n^2)-\frac{k^2n^2(n+1)^2}{4nk}\\
&=\frac{kn(n+1)(2n+1)}{6}-\frac{kn(n+1)^2}{4}\\
&=kn(n+1)\left(\frac{2n+1}{6}-\frac{n+1}{4}\right)\\
&=kn(n+1)(n-1)/12=k(n^3-n)/12,\\
\mathrm{SSR}&=\frac{1}{k}\mathrm{S}=\sum R_{i\cdot}^2/k-\left(\sum R_{i\cdot}\right)^2/nk\\
&=\sum R_{i\cdot}^2/k-k^2n^2(n+1)^2/4nk\\
&=\sum R_{i\cdot}^2/k-kn(n+1)^2/4.
\end{aligned}$$

因此Kendall协和系数W可以表示为

$$\begin{aligned}
W&=\frac{\mathrm{SSR}}{\mathrm{SST}}=\frac{\sum R_{i\cdot}^2/k-kn(n+1)^2/4}{k(n^3-n)/12}\\
&=\frac{\sum R_{i\cdot}^2-k^2n(n+1)^2/4}{k^2(n^3-n)/12}\\
&=\frac{12\mathrm{S}}{k^2n(n^2-1)}.
\end{aligned}\tag{6.7}$$

关于Kendall协和系数检验W的零分布表可以通过下列χ^2公式简单推导得到：

由于

$$\begin{aligned}\mathrm{var}(R_{ij}) &= \frac{\mathrm{SST}}{n-1}\cdot\frac{n}{n-1}\left(\text{以}\frac{n}{n-1}\text{为校正系数}\right)\\ &= \frac{kn(n+1)(n-1)}{12nk}\cdot\frac{n}{n-1}=\frac{n(n+1)}{12},\end{aligned}$$

因此

$$\begin{aligned}\chi^2 &= \frac{\mathrm{SSR}}{\mathrm{var}(R_{ij})}=\frac{\sum R_{i\cdot}^2/k-kn(n+1)^2/4}{\frac{n(n+1)}{12}}\\ &= \frac{\sum R_{i\cdot}^2-k^2n(n+1)^2/4}{kn(n+1)/12}.\end{aligned}$$

由于

$$\begin{aligned}W &= \frac{\sum R_{i\cdot}^2-k^2n(n+1)^2/4}{k^2n(n+1)(n-1)/12}\\ &= \frac{1}{k(n-1)}\frac{\sum R_{i\cdot}^2-k^2n(n+1)^2/4}{kn(n+1)/12}\\ &= \frac{1}{k(n-1)}\chi^2,\end{aligned}$$

因此, Kendall指出, 对于固定的n, 当$k\to\infty$时,

$$k(n-1)W\to\chi^2_{n-1}. \tag{6.8}$$

这样, 对于较大的k, 就可以用极限分布进行检验.

当样本中有结时, 用平均秩方法定秩, 记号不变,

$$\begin{aligned}W_{\mathrm{c}} &= \frac{\sum_{}^{n}R_{i\cdot}^2-\left(\sum R_{i\cdot}\right)^2/n}{\frac{k^2(n^3-n)-k\sum T}{12}}\\ &= \frac{12\sum R_{i\cdot}^2-3k^2n(n+1)^2}{k^2(n^3-n)-k\sum_{}^{g}(\tau_i^3-\tau_i)}.\end{aligned} \tag{6.9}$$

式中,τ_i为结长,g为结的个数.

例6.5 鹈鹕是我国珍稀保护动物, 现测量10只鹈鹕的翼长(X_1)、体长(X_2)及嘴长(X_3)如表6.8所示, 试检验这三组数据是否相关.

表6.8 10只鹈鹕的翼长(X_1)、体长(X_2)及嘴长(X_3)数据表

	翼长(X_1/cm)		体长(X_2/cm)		嘴长(X_3/cm)		
鹈鹕编号	数据	秩	数据	秩	数据	秩	秩和($R_{i\cdot}$)
1	41	7.5	55.7	8	8.6	7.5	23
2	43	9	56.3	9	9.2	9	27
3	39.5	4	54.5	4	8	5.5	13.5
4	38	1	54.2	1.5	5.6	1	3.5
5	40.5	6	55.1	6	6.8	2	14
6	41	7.5	55.4	7	8	5.5	20
7	40	5	54.5	4	8.6	7.5	16.5
8	38.5	2	54.2	1.5	7.4	3.5	7
9	44	10	56.9	10	9.8	10	30
10	39	3	54.5	4	7.4	3.5	10.5
							165

解 假设检验问题为

$$H_0: \text{翼长、体长及嘴长不相关},$$
$$H_1: \text{翼长、体长及嘴长相关}.$$

计算秩统计量如下：

$$\sum R_{i\cdot}^2 - \left(\sum R_{i\cdot}\right)^2/n = 23^2 + \cdots + 10.5^2 - 165^2/10$$
$$= 3380 - 2722.5 = 657.5,$$

$$k^2(n^3-n) = 3^2 \times (10^3 - 10) = 8910,$$
$$\sum T = (2^3-2) + (2^3-2) + (3^3-3) + (2^3-2)$$
$$+(2^3-2) + (2^3-2) = 54.$$

由式(6.9)得

$$W_{\mathrm{c}} = \frac{657.5}{\dfrac{8910 - 3 \times 54}{12}} = \frac{657.5}{729} = 0.9019.$$

由式(6.8), 有

$$\nu = n - 1 = 10 - 1 = 9,$$

$$\chi_\nu^2 = 3 \times (10-1) \times 0.9019 = 24.3513 > \chi_{0.05,9}^2 = 16.9190.$$

由上式χ^2的检验结果, 接受H_1, 翼长、体长及嘴长相关, 呈现一致性.

§6.4 Kappa一致性检验

实际中在做重大决策的时候, 有时需要针对同一研究对象, 进行两组或更多组独立的评判, 如果不同组的结果吻合, 决策更可靠. 反之, 如果两组结果不吻合, 说明决策可能存在着一定的风险, 因而产生了不同组评判结果的一致性检验问题. 这称为结果的一致性问题.

例如, 两家不同医院的专家、医师对同一X光片会诊诊断结果是否相同; 对同一位求职面试者, 假定他经过两个阶段的面试, 前后两阶段的考官组的评分结果是否一致, 或同一研究者, 在不同时间对同一事件的观点是否一致; 等等.

本节仅以两个变量为例, 说明一致性检验的基本原理. 即有假设检验问题:

$$H_0: \text{两种方法不一致} \leftrightarrow \quad H_1: \text{两种方法一致}. \tag{6.10}$$

假设评分是分类或顺序变量, 所有可能的类别为r个. 可以用$r \times r$列联表表示两组结果一致或不一致的频数. 设p_{ij}为对同一事件第一组判为第i类而第二组判为第j类的概率. 若两组判别结果皆相同, 也就是说, 不同专家得到的两组结果完全吻合, 则$p_{ij} = 0, i \neq j$. 而概率和为

$$P_0 = \sum^{r} p_{ii}, \quad r\text{为类别项数}.$$

与一致性结果相反的是独立性, 若各类别的观测值相互独立, 则判断结果皆相同的概率应满足

$$P_{\mathrm{e}} = \sum p_{i\cdot} p_{\cdot i}.$$

式中, $p_{i\cdot}$为第一组专家判为第i类的边缘概率, $p_{\cdot i}$为第二组专家判为第i类的边缘概率, P_{e}表示的是一致性期望概率, 因而$P_0 - P_{\mathrm{e}}$为实际与独立判断结果概率之差. 科恩(Cohen,1960)提出用Kappa统计量表示同一事件, 多次判断结果一致性的度量值如下式所示:

$$K = \frac{P_0 - P_{\mathrm{e}}}{1 - P_{\mathrm{e}}}, \tag{6.11}$$

当$P_0 = 1$时, $K = 1$, 这表示$r \times r$列联表中非对角线上的数据都为0, 一致性非常好. 若$P_0 = P_{\mathrm{e}}$, 即$K = 0$, 则认为一致性较差, 其判断结果完全是由随机产生的独立事件. 另外, K越接近于1, 表示有越高的一致性, 若K接近于0, 则表示一致性较低. 有时K也会有负值, 但很少发生.

经验指出, K的取值与一致性有下表的关系:

表6.9 Kappa经验值

$K < 0.4$	$0.4 < K < 0.8$	$K \geqslant 0.8$
一致性较低	中等一致性	一致性理想

有了估计量, 也可以通过检验判断K值是否为0. 首先计算K的方差如下：

$$\mathrm{var}(K)=\frac{1}{n(1-P_{\mathrm{e}})^2}\left[P_{\mathrm{e}}+P_{\mathrm{e}}^2-\sum p_{i\cdot}p_{\cdot i}(p_{i\cdot}+p_{\cdot i})\right] \tag{6.12}$$

科恩（Cohen）于1960年指出, K在大样本下有正态近似：

$$Z=\frac{K}{\sqrt{\mathrm{var}(K)}}, \tag{6.13}$$

如果$Z>Z_{0.05/2}=1.96$, 则表示$K>0$, 表示有一致性.

例6.6 假设某啤酒大赛中, 多种品牌的啤酒由来自甲、乙两地的专业品酒师进行评分, 每个品牌只允许选送一种酒作为代表参评, 每位品酒师对每种啤酒将按照3个级别评分, 结果如表6.10所示, 其中第i,j位置的n_{ij}表示甲评分为i, 而乙评分为j的累积品牌数.

按式(6.11)计算概率：

表6.10 两组品酒师评分频数交叉列联表

		乙地 (级别)			行和
甲地		1	2	3	
级	1	18(0.36)	2(0.04)	0(0)	20(0.40)
	2	4(0.08)	12(0.24)	1(0.02)	17(0.34)
别	3	2(0.04)	1(0.02)	10(0.20)	13(0.26)
列和		24(0.48)	15(0.30)	11(0.22)	50(1.00)

$$\begin{aligned}
P_0&=0.36+0.24+0.20=0.80,\\
P_{\mathrm{e}}&=0.4\times0.48+0.34\times0.30+0.26\times0.22=0.3512,\\
K&=\frac{0.80-0.3512}{1-0.3512}=\frac{0.4488}{0.6488}=0.6917.
\end{aligned}$$

由式(6.12)及式(6.13), 得

$$\begin{aligned}
\mathrm{var}(K)&=\frac{1}{50(1-0.3512)^2}\{0.3512+0.3512^2\\
&\quad-[0.4\times0.48(0.4+0.48)+0.34\times0.3(0.34+0.3)\\
&\quad+0.26\times0.22(0.26+0.22)]\}\\
&=\frac{0.2128454}{21.04707}=0.0101128,\\
\sqrt{\mathrm{var}(K)}&=\sqrt{0.0101128}=0.1005624,\\
Z&=\frac{0.6917}{0.1005624}=6.8783>Z_{0.05/2}=1.96.
\end{aligned}$$

因此一致性不为0, 而$K=0.6917$, 表示甲、乙两地品酒师的评分保持较好的一致性.

§6.5 HBR基于秩的稳健回归

在第4章的方差分析中,当正态性条件的前提不能获得满足时，就需要引入基于观测的秩统计量建立非参数检验。如果回归分析中的残差项不满足正态性假设，比如有离群值存在的时候。很自然就会将基于秩的想法扩展到对误差的分析中. 最早是雅克尔(L.A.Jeackel)(1972)和德雷珀(Draper)(1998)等多位学者提出了基于秩的R估计法，将秩的某个得分函数作为权重引入估计模型用以降低离群点的不良影响. 之后，为提高R估计稳健性，张(Chang)(1999)提出HBR高失效点(High Breakdown Point)的R估计。这一节将主要介绍基于残差秩的稳健回归方法中的参数R稳健估计、稳健性质和回归诊断。

§6.5.1 基于秩的R估计

1.R估计函数.

假设有回归模型$y_i = x_i\beta + r_i, i = 1, \cdots, n$,其中$r_i = y_i - x_i\beta$为第$i$个样本的残差，$R(r_i)$为第$i$个残差的秩，$a(R(r_i))$为残差秩的得分函数，定义R估计得分为：

$$D_{\mathrm{R}}(\hat{\beta}) = \sum a(R(r_i))r_i,$$

得分函数$a(i) = \phi(\frac{i}{n+1})$.其中最常用的是Wicoxon 得分函数：$\phi(u) = \sqrt{12}(u-1/2)$.带入上面的定义，得到该估计的目标函数为：

$$D_{\mathrm{R}}(\hat{\beta}) = \frac{\sqrt{12}}{n+1}\sum_{i=1}^{n}\left(R(r_i) - \frac{n+1}{2}\right)r_i$$

对其求极小值，得到相应的偏回归系数的Wilcoxon R估计值：

$$\hat{\beta}_{\mathrm{R}} = \mathrm{argmin}||y - x\beta||_{\mathrm{HBR}} = \mathrm{argmin}D_{\mathrm{R}}(\beta)$$

2.GR估计函数.

$D_{\mathrm{GR}}(\beta) = ||y - x\beta||_{\mathrm{GR}} = ||u||_{\mathrm{GR}} = \sum\sum_{i<j} b_{ij}|u_i - u_j|$, 其中$u_i$为残差，$b_{ij}$为正的对称权重，$b_{ij} = b_{ji}$, 当$b_{ij} \equiv 1$时，该方程退化为前面的Wilcoxon R估计量，这时即有$D(\beta) = \sum\sum_{i<j}|u_i - u_j| = 2\sum_{i=1}^{n}(R(u_i) - (n+1)/2)u_i$. 选择合适的$b_{ij}$作为权重函数可用来减小$X$空间离群点的影响，一般情况下，$b_{ij}$的定义如下：

$$b_i = \min\{1, c_1/\sqrt{h_{ii}}\}^{\alpha_1}, \quad b_{ij} = b_i \times b_j,$$

其中，$\alpha_1 = 2, h_{ii}$为观测点i的杠杆值，定义为帽子矩阵$H = X(X^TX)^{-1}X^T$的主对角线第i 位置上的元素。表示X方向上该点距离中心位置的远近，c_1一般取杠杆值0.70分位数。从这些式子中可以观察到，某观测值的杠杆值越大，该点位置距离中

心位置越远，在权值函数b_{ij}中的取值越小，离群值在X空间中对β 的估计影响越小。不过，当X空间存在多个离群值时，杠杆值的计算比较敏感，基于杠杆值的权重函数稳健性不佳，这时可以考虑使用马氏距离定义$MCD_i = (x_i - \bar{X})^T(X^TX)^{-1}(x_i - \bar{X})$,可以得到$MCD_i$与杠杆值$h_i$ 存在如下关系：

$$h_i = \frac{MCD_i}{n-1} + \frac{1}{n}.$$

对于普通的马氏距离，由于X空间极端的多个离群值可使MCD_i中的均值和协方差阵发生较大偏离，MCD_i和杠杆值一样不具有稳健性，但马氏距离中的均值和协方差是可以做稳健化处理的，这样就可以得到广义的$Mallows$(Generalized Mallows Weights)权重:

$$b_i = \min\{1, c_2/(x_i - v)^TV^{-1}(x_i - v))^{1/2}\}^{\alpha_2},$$

其中，c_2可以取自由度为p的χ^2分布的0.95分位数，$\alpha_2 = 2, (v, V)$是位置和离散程度MVE(minimum volume ellipsoid)或MCD(minimum covariance determinant)估计量，前者的估计值反映了一半数据中的最小置信域体积，后者从包含一般数据最小协方差阵行列式得到，求解的方法一般采用重复抽样算法，具体的计算可参见文献伍德拉夫(Woodruff D.L.)(1993)。从GR 估计函数中可以看出其中的各元素同时具有对X 空间距离和残差的降权作用，而R估计仅仅是对残差做了降权，对GR 估计函数求极小化，可得到参数的GR 估计，数值解法中R 和GR 均可采用梯度法实现，而且满足位置和尺度同变性。

3.HBR估计函数.

定义

$$D_{\text{HBR}}(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} a_n(R_{ni}^+)|r_i|$$

r_i如上是第i个样本点的残差，R_{ni}^+是满足条件的残差绝对值的秩，$a_n(i)$为得分函数，如$a_n(i) = h^+(i/(n+1))$.h^+的选择借用了高失效点LTS（Least Trimmed Sum of Squares）回归的想法，它不仅考虑了X 空间中的每个点的权值，还考虑了Y空间的权值。例如，只选择残差排序在较小前$\alpha(0 < \alpha < 1)$ 的观测，此时$h^+ \approx n \times \alpha$, 这样就会令残差绝对值排序在较大的$1-\alpha$部分不必用于参与$D_{\text{HBR}}$的计算，可以得到该估计函数失效点为$\epsilon^* = \min\{\alpha, 1-\alpha\}$.（详见参考文献罗素(Rousseeuw) 和范德丽森(Van Driessen)(1999)）当$\alpha = 0.5$ 时，最大的失效点为50%,在实际应用中的失效点可以通过α来控制失效点，也可以通过重复抽样来获得。

§6.5.2 假设检验

假设要对p个回归估计中的q个参数做假设检验，两种R估计都可以用F检验统

计量

$$F_{\mathrm{R}} = \frac{RD_{\mathrm{R}}/q}{\hat{\tau}_R^2}$$

$$F_{\mathrm{GR}} = \frac{\sqrt{12}}{n}\frac{RD_{\mathrm{GR}}/q}{\hat{\tau}_R}$$

其中$RD_\phi = D_\phi(Y,p) - D_\phi(Y,p-q)$,表示缩减模型(包含$p-q$个估计参数)与完全模型（包含$p$个参数）之间的离差函数的减小量(reduction in residual dispersion),$\hat{\tau}$ 类似于LS估计中的剩余标准差。

§6.5.3 多重决定系数CMD(Coefficient of Multiple Determination)

$$R_R^2 = \frac{RD}{RD + (n-p-1)(\hat{\tau}/2)}$$

$$R_{GR}^2 = \frac{RD_{\mathrm{GR}}}{D_{\mathrm{GR}} + \frac{n(n-p-1)}{\sqrt{12}}\hat{\tau}}$$

在回归分析中，CMD是一个反映拟合效果的统计量，当完全拟合时，取值为1，完全失拟时，取值为0，它能较好地反应模型拟合的效果，且由于它与稳健的假设检验统计量F相联系，该统计量是稳健的。

§6.5.4 回归诊断

1.残差图.

R和HBR估计的残差图与LS估计类似，例如用残差与预测值作图得到的图形的分布不是围绕0 随机波动而是出现某种曲线趋势，则此残差图提示拟合所用的模型假设不恰当，但是由于GR估计的拟合值和残差均是权重的函数，因此在解释该残差图时就不像R-LS残差图那样意义明确，但均可用于初步识别可能的离群值。

2.标准化残差.

三种估计方法都可用各自残差除以其残差标准误的估计值来得到标准化残差，当标准差残差绝对值大于3 的时候，判为潜在的离群值。

3.影响数据的度量. R估计中使用量RFIT某个数据点对模型拟合造成的影响，定义为：

$$RFIT = \hat{Y}_{R_i} - \hat{Y}_{R(-i)},$$

其中$\hat{Y}_{R_i}$指第i个数据点的拟合值，$\hat{Y}_{R(-i)}$为删除第i个数据点的模型Y_i 的预测值，GR估计的影响诊断中，首先需要计算诊断量TAS_R 来比较对同一份资料的进行R 估计和GR 估计时，两种方法的整体差异，计算公式为：

$$TAS_R = (\hat{b}_{\mathrm{R}} - \hat{B}_{\mathrm{GR}})^T A_{\mathrm{R}}^{-1}(\hat{b}_{\mathrm{R}} - \hat{B}_{\mathrm{GR}})$$

其中

$$A_R = \begin{pmatrix} \hat{\alpha}_R \\ \hat{\beta}_R \end{pmatrix} = \begin{pmatrix} \hat{\tau}^2 & 0 \\ \hat{0} & \tau(X^TX)^{-1} \end{pmatrix} \tag{6.14}$$

该统计量界值为$\frac{4(p+1)^2}{n}$.对同一数据的R估计与GR估计拟合的总体差异是由于对高杠杆值进行降权处理而引起的，如果该值较大则需要计算如下$CS_{R,i}$诊断量：

$$CS_{R,i} = \frac{\hat{y}_{R,i} - \hat{y}_{GR_i}}{(n^{-1}\hat{\tau}^2 + h_i\hat{\tau}^2)^{1/2}}$$

该诊断量用以识别那个点对两估计方法差异的贡献大，其界值为$2\sqrt{(n-p)/n}$.

例6.7 该数据来自Rousseeuw(1987),是有关CYG OB1星团的天文观测数据，该星团包含47颗恒星，对每颗星球的发光强度和球面温度进行测量。响应变量为对数光强（用light表示），解释变量为对数温度（用temperature表示），如图6.1 制作了这两个变量的散点图，其中有4 颗恒星光强度异常，温度较低，光强度却和星团其他成员强度相当。在该数据集中，这4 颗恒星被标记为巨星，另外还有两颗恒星对数温度分别为3.84 和4.01，除了这6颗星以外，其他41颗星被标记为该星团的主序星。我们尝试了三种回归估计目标函数，分别为最小二乘(LS)，GR估计，Wilcoxonϕ函数以及HBR稳健估计回归。

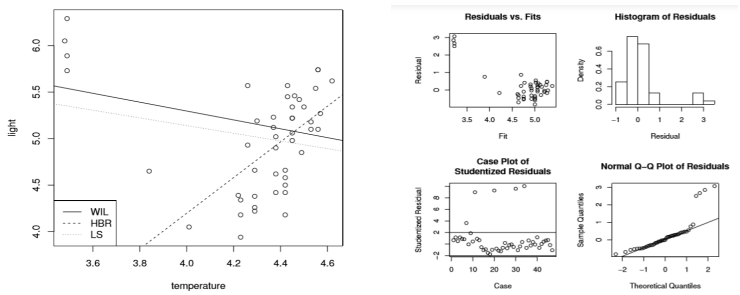

图 6.1 恒星表面温度和发光强度散点图、三种估计拟合线及HBR估计诊断图

该分析程序如下：

```
data(stars)
fitHBR<-hbrfit(stars$light ~ stars$temperature)
diagplot(fitHBR)#制作诊断图
> summary(fitHBR)
Call: hbrfit(formula = stars$light ~ stars$temperature)
Coefficients:
```

```
                  Estimate  Std. Error  t.value p.value
(Intercept)      -3.46917    1.64733   -2.1059 0.04082 *
stars$temperature 1.91667    0.38144    5.0248 8.47e-06***
--Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Wald Test: 25.24853 p-value: 1e-05
```

由图6.1中的诊断图来看,HBR不仅可以识别出4颗强度较大的恒星，而且可以识别出两个温度较小的异常恒星，而且它的回归线穿过了41 颗主序恒星，而另外两种方法建立的回归线明显受到两类异常的误导，表现出了较差的拟合效果。

§6.6 中位数回归系数估计法

回归分析是统计学中应用最广泛的方法之一. 回归分析主要是刻画变量和变量之间的依赖关系. 一个简单的一元回归模型如下定义：给定数据点$(X_i, Y_i), i = 1, 2, \cdots, n$, 假定$Y$的平均变动由$X$决定, 那么不能由$X$解释的部分用噪声$\varepsilon$表示. Y与X的关系表示如下：

$$Y_i = \alpha + \beta X_i + \varepsilon_i,\ i = 1, 2, \cdots, n, \tag{6.15}$$

其中α, β是待估的未知参数, ε_i为来自某未知分布函数$F(x)$的误差. ε_i一般要满足Gauss-Markov 假设条件, 即

$$\begin{aligned} &E\varepsilon_i = 0, \quad i = 1, 2, \cdots, n. \\ &\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j; \\ 0, & i \neq j, i, j = 1, 2, \cdots, n. \end{cases} \end{aligned} \tag{6.16}$$

实际中许多问题不满足诸如此类的假设条件, 比如等方差假设就很难满足,最小二乘法估计回归系数的方法受到了挑战, 结果就产生了非参数系数估计的方法.这里我们介绍两种基于秩的非参数系数估计法——Brown-Mood 法和Theil 法.

§6.6.1 Brown-Mood方法

该方法是由布朗(Brown)和慕德(Mood)于1951年在一次会议中提出的. 为了估计α和β, 首先找到X 的中位数X_{med}, 将数据按照X_i是否小于X_{med}分成两组 Ⅰ, Ⅱ, 第 Ⅰ 组数据中$X_i < X_{\text{med}}$, 第 Ⅱ 组数据中$X_i > X_{\text{med}}$; 然后, 在两组数据中分别找到两个代表值, 令$X'_{\text{med}}, Y'_{\text{med}}$ 分别是第 Ⅰ 组样本的中位数, $X''_{\text{med}}, Y''_{\text{med}}$分别是第 Ⅱ 组样本的中位数, β的估计值为

$$\hat{\beta}_{BM} = \frac{Y'_{\text{med}} - Y''_{\text{med}}}{X'_{\text{med}} - X''_{\text{med}}}, \tag{6.17}$$

这个估计值是回归直线的斜率的估计. 因而, 回归直线在Y轴上的截距α的估计值为

$$\hat{\alpha}_{BM} = \text{median}\{Y_i - \hat{\beta}_{BM}X_i, i = 1, 2, \cdots, n\}.$$

例6.8 参见南非心脏病数据中的ldl(低密度脂蛋白),adiposity(肥胖指标), 这两项指标之间存在着一定的关系.首先通过画这15 个点的散点图(见图6.2)可以看出, ldl(低密度脂蛋白)增大, adiposity(肥胖指标)有增加趋势.我们编写如下程序计算中位数回归直线:

```
yy=adiposity
xx=ldl

cyx=coef(lm(yy~xx))

md=median(xx)
xx1=xx[xx<=md]
xx2=xx[xx>md]
yy1=yy[xx<=md]
yy2=yy[xx>md]
md1=median(xx1)
md2=median(xx2)
mw1=median(yy1)
mw2=median(yy2)
beta=(mw2-mw1)/(md2-md1)
alpha=median(yy-beta*xx)
plot(xx,yy)
abline(alpha,beta)
abline(c(cyx),lty=2)
title("Brown-Mood median regression")
```

由计算公式得$\hat{\beta}_{BM} = 2.9523, \hat{\alpha}_{BM} = 11.5552$,于是所求中位数回归直线为: adiposity $=$ 2.9523ldl$+$11.5552. 图中的长虚线是最小二乘估计, 回归方程为adiposity $=$ 1.65ldl$+$ 17.562,从图中看, 最小二乘估计显然偏离了主体数据的走向, 原因是它较易受到异常数据的拉动影响.

§6.6.2 Theil方法

6.6.1小节介绍的Brown-Mood方法估计回归系数的方法较为粗糙, 它只用到样本中位数的信息, 没有用到样本中更多的信息. 与之相比, 泰尔(Theil)于1950年提

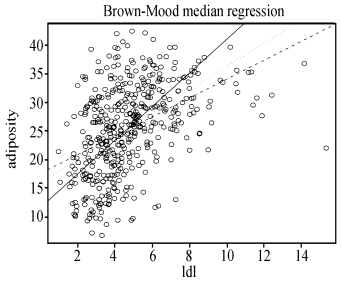

图 6.2 Brown-Mood、Theil中位数和最小二乘回归直线图

出的Theil方法则将Brown-Mood方法发展到所有的样本上. 他的基本原理在于, 对于任意两个横坐标不相等的点, 如$(X_i, Y_i), (X_j, Y_j)$, 根据斜率β的几何意义, 可以用$\dfrac{Y_j - Y_i}{X_j - X_i}$估计一个$\beta_{ij}$, 将所有斜率平均则可以作为$\beta$的估计值,于是有了下面的估计.

假设自变量X中没有重复数据, 任给$i < j$, 记$s_{ij} = \dfrac{Y_j - Y_i}{X_j - X_i}$, 则$\beta$ 的估计为

$$\tilde{\beta}_T = \text{median}\{s_{ij} : 1 \leqslant i < j \leqslant n\}.$$

相应地, α的估计值取为

$$\tilde{\alpha}_T = \text{median}\{Y_j - \tilde{\beta}_T X_j : j = 1, 2, \cdots, n\}.$$

当自变量X中有相等数据存在时, 如$(X_1, Y_1), \cdots, (X_1, Y_l)$. 记$Y^* = \text{median}\{Y_i : 1 \leqslant i \leqslant l\}$, 也就是说用一个点$(X_1, Y^*)$代替上面的$l$个样本点后, 再用无节点方法计算$\tilde{\alpha}$和$\tilde{\beta}$ 即可.

例6.9(例6.8续) 对于例6.8中的数据, 用Theil方法重新计算β_T和α_T的估计值为$\tilde{\beta}_T = 2.029, \tilde{\alpha}_T = 16.024$, 于是回归直线为adiposity $= 16.024 + 2.029$ldl. 我们发现Theil方法得到的趋势线界于Brown-Mood方法和最小二乘回归直线之间.

§6.6.3 关于α和β的检验

关于α和β的假设检验, 我们感兴趣的假设检验可能有如下两种：

$$H_0 : \alpha = \alpha_0 \ \beta = \beta_0 \leftrightarrow H_1 : \alpha \neq \alpha_0 \ \text{ 或 } \ \beta \neq \beta_0. \tag{6.18}$$

$$H_0' : \beta = \beta_0 \leftrightarrow H_1' : \beta \neq \beta_0. \tag{6.19}$$

对于第一种假设问题$H_0 \leftrightarrow H_1$, 主要判断回归直线是否比较均衡地反映了数据的分布, 我们介绍Brown-Mood检验作为代表; 对于第二种假设问题, 以Theil检验为代表. 无论哪一种检验, 对上面两种方法都适用.

1. Brown-Mood检验

对于假设问题(6.18), 在H_0下回归直线为$y = \alpha_0 + \beta_0 x$, 如果回归直线比较理想, 则所有的数据点应该比较均匀地分布在回归直线的上下两侧, 也就是说, 回归直线上下两侧(X_i, Y_i)的个数应比较接近$\frac{n}{2}$. 仅仅如此还是不够的, 如果回归直线左右样本点不均衡, 比如较大的自变量更倾向于在回归直线的下方, 而另一侧则堆积了更多自变量较小的样本点, 那么就表示回归直线不理想. 于是可以用回归直线的左上与右下的样本点个数是否相等来衡量原假设H_0.

具体而言, 记$X_{\text{med}} = \text{median}\{X_1, X_2, \cdots, X_n\}$,

$$\begin{aligned} n_1 &= \#\{(X_i, Y_i) : X_i < X_{\text{med}},\ \ Y_i > \alpha_0 + \beta_0 X_i\}, \\ n_2 &= \#\{(X_i, Y_i) : X_i > X_{\text{med}},\ \ Y_i < \alpha_0 + \beta_0 X_i\}. \end{aligned}$$

由上面分析可知, 当H_0成立时, $n_1 \approx n_2 \approx \frac{n}{4}$; 而当$H_1$成立时, n_1 与n_2 中至少有一个远离$\frac{n}{4}$. 于是我们可以用

$$\left(n_1 - \frac{n}{4}\right)^2 + \left(n_2 - \frac{n}{4}\right)^2$$

作为检验统计量, 其取大值时拒绝H_0.

为了大样本近似的方便, Brown和Mood于1951年提出用

$$\text{BM} = \frac{8}{n}\left[\left(n_1 - \frac{n}{4}\right)^2 + \left(n_2 - \frac{n}{4}\right)^2\right]$$

作为检验统计量, 称为Brown-Mood检验统计量. 当BM取大值时拒绝H_0. 关于Brown-Mood检验的零分布表, 没有现成表可用. 但是, 二人于1950年还证明, 当$n \to \infty$时, 有

$$\text{BM} \longrightarrow \chi^2(2).$$

类似于关于$H_0 \leftrightarrow H_1$的Brown-Mood检验统计量的得出, Brown和Mood在同一篇文章中提出, 关于假设$H_0' \leftrightarrow H_1'$, 可以用统计量

$$\text{BM}' = \frac{16}{n}\left(n_1 - \frac{n}{4}\right)^2$$

检验, 其中

$$n_1 = \#\{(X_i, Y_i) :\ \ X_i < X_{\text{med}},\ \ Y_i > a + \beta_0 X_i\}.$$

H_0的拒绝域为其取大值. 我们也称为Brown-Mood检验.

另外, Brown和Mood证明, $n \to \infty$时, 有

$$\mathrm{BM}' \to \chi^2(1).$$

例6.10(例6.8续) 对于例6.8中的数据, 通过Brown-Mood估计, 估计回归直线为$y = 2.9523x + 11.5552$. 以下用Brown-Mood检验对回归直线的均衡性进行检验, 即要检验

$$H_0 : \alpha = 11.5552, \quad \beta = 2.9523,$$

$n_1 = 126, \ \ n_2 = 104$. 经计算得

$$\mathrm{BM} = \frac{8}{462}\left[\left(126 - \frac{462}{4}\right)^2 + \left(104 - \frac{462}{4}\right)^2\right] \approx 4.1991.$$

双边检验p值为0.12, 因而没有理由拒绝H_0, 没有违背均衡性. 如果将同样的过程应用于OLS回归直线$y = 1.6548x + 17.5626$, 计算得$\mathrm{BM} = 6.7965$, 双边检验p值为$0.033 < 0.1$, 认为回归直线违背均衡性,这一结论与图形观察结果是一致的. Theil方法建立的回归方程的均衡性检验留作习题.

2. Theil检验

对于假设问题(6.18), 还有一种基于Kendall τ检验和Spearman秩相关系数给出的处理方法. 我们注意到, 当回归直线$y = \alpha + \beta_0 x$拟合数据$(X_1, Y_1), \cdots, (X_n, Y_n)$较好时, 说明$Y_i - \beta_0 X_i$只受一个系统因素$\alpha$和随机误差的影响, 而与自变量$X_i$没有什么关系, 于是我们可以用$X_i$与$Y_i - \beta_0 X_i$相关与否衡量$H_0' : \beta = \beta_0$. 如果相关性很大, 则认为假设检验$H_0' : \beta = \beta_0$不成立, 测量相关性, 我们讲过可以用Kendall τ和Spearman秩相关系数等检验. 这样, Theil于1950年提出用基于Kendall τ的方法来检验$H_0' \leftrightarrow H_1'$. 只是注意到, 此时的$R_i, Q_i$分别表示$X_i, Y_i - \beta_0 X_i$在$(X_1, \cdots, X_n)$和$(Y_1 - \beta_0 X_1, \cdots, Y_n - \beta_0 X_n)$中的秩或者平均秩, 故我们称为Theil检验.

例6.11(例6.8续) 对于例6.8中的数据, 如果用Theil检验, Theil中位数回归假设$H_0 : \beta = 2.029$如下：利用Theil回归的β的估计(即$\beta_T = 2.029$), 得到一系列残差reyTH= $\{e_i\}$. 相应的关于$(x_1, e_1), \cdots, (x_{25}, e_{25})$的Kendall相关系数计算如以下程序：

```
cor.test(reyTH,x,me="kendall")
        Kendall's rank correlation tau
data:  reyTH and xx
z = -0.0051, p-value = 0.996
alternative hypothesis: true tau is not equal to 0
```

```
sample estimates:
        tau
-0.000159776
```

Kendall相关系数$\tau = 0.00016$.零假设下的p值为0.996, 双边检验协同意义下, 没有理由拒绝零假设$H_0 : \beta = 2.029$这个回归系数.

§6.7 线性分位回归模型

分位回归(quantile regression)是由科恩克(Koenker)和巴塞特(Bassett)于1978年提出的, 其基本思想是建立因变量Y对自变量X的条件分位数回归拟合模型, 即

$$Q_Y(\tau|X) = f(X),$$

其中, τ是因变量Y在X条件下的分位数. $f(X)$拟合Y的第τ分位数, 于是中位数回归就是0.5分位回归. 如果将τ从$0.1, 0.2, \cdots, 0.9$取值, 就可以解出9个回归方程.

传统的回归建立在假设因变量Y和自变量X有如下关系的基础上：

$$E(Y|X) = f(X) + \epsilon.$$

对任意的$X = x$,当ϵ满足正态和齐性(方差相等)条件时, 可以用最小二乘法建立回归预测模型. 实际情况下, 这两个假设往往得不到满足, 比如ϵ左偏或右偏, 用最小二乘拟合回归模型稳定性很差. 分位回归对分位数进行回归, 不需要分布和齐性方面过强的假设, 在ϵ非正态和非齐性的情况下也能较好地把握数据的主要规律. 分位回归以其稳健的性质已经开始在经济和医学领域广泛应用, Koenker和哈洛克(Hallock)(2001)给出了这方面的很多应用. 本节我们着重介绍线性分位回归模型及应用.

已知观测$(\boldsymbol{X}, Y) = \{(\boldsymbol{x}_i, y_i), i = 1, 2, \cdots, n, y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^p\}$. X对Y的线性分位回归模型为

$$Q_Y(\tau|\boldsymbol{X}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}. \tag{6.20}$$

怎样求解其参数? 线性回归通过最小化残差平方和求解, 中位数回归通过最小化残差的绝对值求解, 显然线性分位回归可以通过最小化残差绝对值加权求和, 只是在绝对值前应增加分位点权重系数即可. 于是线性分位回归的最优化问题表示为

$$\hat{\beta} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}). \tag{6.21}$$

式中, ρ_τ是权重函数, 表示实际值与拟合值位置关系的权重比例. τ分位回归中小于分位点的可能性为τ, τ分位回归中不小于分位点的可能性为$1-\tau$. ρ_τ如下理解:

$$\rho_\tau(u)=\begin{cases}\tau u, & u\geqslant 0,\\ (1-\tau)|u|, & u<0.\end{cases}$$

给定τ,注意到式(6.21)等价于

$$\hat{\beta}(\tau)=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left[\sum_{i\in\{i:y_i\geqslant \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)\}}\tau|y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)|_+ + \sum_{i\in\{i:y_i< \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)\}}(1-\tau)|y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)|_-\right].$$

Koenker和奥利(Orey)(1993)运用运筹学中的单纯形法求解线性分位回归, 其思想是: 任选一个顶点, 沿着可行解围成的多边形边界搜索, 直到找到最优点. 该算法估计出来的参数具有很好的稳定性,但是在处理大型数据时运算的速度会显著降低. 目前流行的还有内点算法(interior point method)和平滑算法(smoothing method)等. 由于分位回归需要借助大量计算, 模型的参数估计要比传统的线性回归模型的求解复杂.

除参数回归模型、分位回归模型外, 还有非参数回归模型、半参数回归模型等, 不同的模型都有相应的估计方法.

与线性最小二乘回归相比较, 分位回归的优点体现在以下几方面:

(1) 分位回归对模型中的随机误差项不需对分布做具体的假定, 有广泛的适用性;

(2) 分位回归没有使用连接函数描述因变量与自变量的相互关系, 因此分位回归体现了数据驱动的建模思想;

(3) 分位回归对分位数τ进行回归, 于是对于异常值不敏感, 模型结果比较稳定;

(4) 由分位回归解出的系列回归模型可更为全面地体现分布特点.

例6.12 这是Koenker给出的一个例题, Engel Data(恩格尔数据), 研究者对235个比利时家庭的当年家庭收入(income)和 当年家庭用于食品支出的费用(foodexp)进行观测. 在R中用分位回归建立恩格尔数据的等间隔分位回归. R参考程序如下:

```
install.packages("quantreg") ;library(quantreg); library(SparseM)
par(mfrow=c(1,3)); data(engel); attach(engel);
plot(income,foodexp,xlab="Household Income", ylab="FoodExpenditure",type = "n", cex=.5)
points(income,foodexp,cex=.5);
taus=seq(0.1,0.9,0.1); f=coef(rq((foodexp)~(income),tau=taus));
for(i in 1:length(taus)){
```

```
        abline(f[,i][1],f[,i][2],lty=2)
}
abline(lm(foodexp ~ income),lty=9)
abline(rq(foodexp~income,tau=0.5))
legend(3000,700,c("mean","median","otherquantile"),lty = c(9,1,2))
plot(taus,f[1,]);
lines(taus,f[1,]),plot(taus,f[2,]);lines(taus,f[2,]);
```

图6.3中,从下至上虚线分别为分位数回归($\tau = 0.1, \cdots, 0.9$), 分位数间隔0.1, 实线为最小二乘回归. 注意到, 家庭食品支出随家庭收入增长而呈现增长趋势. 不同τ值的分位回归直线从上至下的间隙先窄后宽说明了食品支出是左偏的, 这一点从分位数系数随分位数增加变化图(最右侧的点)中也可以得到验证. 即在固定收入的时候, 家庭支出密集在较高的位置, 少数家庭支出偏低. 中位数回归直线始终位于最小二乘回归直线之上, 截距显著不同, 说明最小二乘回归显然受到两个异常点(高家庭收入低食品支出) 的 影响较大, 这种不稳定的结果, 就是对贫穷家庭的平均家庭收入预测较差,高估了他们的生活质量.

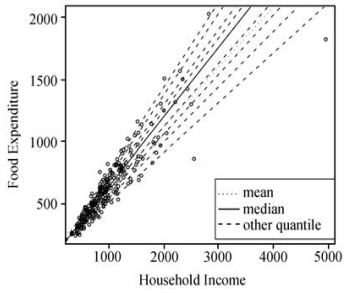

图 6.3 恩格尔数据分位回归

习题

6.1 从中国30个省区抽样的文盲率(单位:‰)和各省人均GDP(单位:元)的数据如下:

文盲率	7.33	10.80	15.60	8.86	9.70	18.52	17.71	21.24	23.20	14.24
人均GDP	15044	12270	5345	7730	22275	8447	9455	8136	6834	9513
文盲率	13.82	17.97	10.00	10.15	17.05	10.94	20.97	16.40	16.59	17.40
人均GDP	4081	5500	5163	4220	4259	6468	3881	3715	4032	5122
文盲率	14.12	18.99	30.18	28.48	61.13	21.00	32.88	42.14	25.02	14.65
人均GDP	4130	3763	2093	3715	2732	3313	2901	3748	3731	5167

运用Pearson, Spearman 和Kendall检验统计量检验文盲率和人均GDP之间是否相关, 是正相关还是负相关.

6.2 某公司销售一种特殊的化妆用品, 该公司观测了15个城市在某季度对该化妆品的销售量Y(单位：万件)和该地区的人均收入X(单位：百元), 如表7.8所示.

序号	1	2	3	4	5	6	7	8
地区X	9.1	8.3	7.2	7.5	6.3	5.8	7.6	8.1
人口Y	8.7	9.6	6.1	8.4	6.8	5.5	7.1	8.0
序号	9	10	11	12	13	14	15	
地区X	7.0	7.3	6.5	6.9	8.2	6.8	5.5	
地区Y	6.6	7.9	7.6	7.8	9.0	7.0	6.3	

以往的经验表明, 销售量与人均收入之间存在线性关系, 试写出由人均收入解释销售量的中位数线性回归直线.

6.3 在歌手大奖赛中, 裁判是根据歌手的演唱进行打分的, 但有时也可能带有某种主观色彩. 此时作为大赛公证人员有必要对裁判的打分是否一致进行检验, 如果一致, 则说明裁判组的综合专家评判的结果是可靠的. 下面是1986年全国第二届青年歌手电视大奖赛业余组民族唱法决赛成绩的统计表, 试进行一致性检验.

	歌 手 成 绩									
裁判	1	2	3	4	5	6	7	8	9	10
1	9.15	9.00	9.17	9.03	9.16	9.04	9.35	9.02	9.10	9.20
2	9.28	9.30	9.31	8.80	9.15	9.00	9.28	9.29	9.10	9.30
3	9.18	8.95	9.24	8.93	9.17	8.85	9.28	9.05	9.10	9.20
4	9.12	9.32	8.83	8.86	9.31	8.81	9.38	9.16	9.17	9.10
5	9.15	9.20	8.80	9.17	9.18	9.00	9.45	9.15	9.40	9.35
6	9.35	8.92	8.91	8.93	9.12	9.25	9.45	9.21	8.98	9.18
7	9.30	9.15	9.10	9.05	9.15	9.15	9.40	9.30	9.10	9.20
8	9.15	9.01	9.28	9.21	9.18	9.19	9.29	8.91	9.14	9.12
9	9.21	8.90	9.05	9.15	9.00	9.18	9.35	9.21	9.17	9.24
10	9.24	9.02	9.20	8.90	9.05	9.15	9.32	9.28	9.06	9.05
11	9.21	9.23	9.20	9.21	9.24	9.24	9.30	9.20	9.22	9.30
12	9.07	9.20	9.29	9.05	9.15	9.32	9.24	9.21	9.29	9.29

6.4 100名牙疾患者, 先后经过两位不同的牙医的诊治, 两位牙医在是否需要进行某项处理时给出的诊治方案不完全一致. 现将两位牙医的不同意见数据列表如下, 试分析两位医生的治疗方案是否完全一致.

		牙医乙		
		需要处理	不需要处理	合计
牙医甲	需要处理	40	5	45
	不需要处理	25	30	55
	合计	65	35	100

6.5 为测量某种材料的保温性能, 把用其覆盖的容器从室内移到温度为x的室外, 三小时后记录其内部温度y. 经过若干次试验, 产生如下记录(单位:华氏度). 该容器放到室外前的内部温度是一样的.

x	33	45	30	20	39	34	34	21	27	38	30
y	76	103	69	50	86	85	74	58	62	88	210

试用Theil和Brown-Mood方法作线性回归.两个线性方程是否一致, 是否存在离群点? 如果存在, 请指出, 并删除它后重新拟合.

6.6 用Brown-Mood方法检验用Theil方法建立的回归方程的均衡性.

6.7 检验例6.9中用Theil法估计得到的回归系数.

6.8 有关分位回归, 回答以下问题.

(1) 简述分位回归模型.

(2) 简述分位回归模型参数估计的最优化问题.

(3) 分位回归相比于线性回归的优点有哪些? 为什么具备这些优点?

(4) 用分位回归方法拟合光盘中的infant-birthweight数据, 并进行解释.

6.9 模拟实验分析：(X,Z)的真实关系满足$z=2\cdot(\exp(-30\cdot(x-0.25)^2)+\sin(\pi x^2))$.从均匀分布$U(0,1)$ 中抽取100 个X 值，将这些数值从小到大排序，依次产生带有$N(0,1)$噪声的Y值,即：$y=z+N(0,1)$.这样的实验重复20 次，得到(X,Y) 观测值矩阵和真值矩阵(X,Z)，完成以下分析任务：

(1) 绘制(X,Y)的散点图，并在散点图上添加由(X,Z)生成的真实函数曲线;

(2) 求解中位数线性回归，0.25分位数线性回归和0.75 分位数线性回归，和不带噪声的真实值进行比较，估计拟合的均方误差;

(3) 将线性回归改为多项式为二阶表示型(模型中纳入X^2项)和四阶(模型中纳入X^2,X^3,X^4项)，继续拟合数据，比较(2) 和(3) 拟合的结果有怎样的不同；

(4) 改变Y值的生成方式：$y=2\cdot(\exp(-30\cdot(x-0.25)^2)+\sin(\pi x^2))+N(0,(2x)^2)$ 求解多项式为二阶(X^2)和四阶(X^2,X^3,X^4)的中位数、0.25分位数、0.75分位回归.将这些拟合线绘制到散点图上。比较(2)(3)(4)的数据分析，给出讨论。

案例与讨论：中医与西医治疗方法之间的差异分析？

案例背景

大肠癌是最常见的恶性肿瘤，老年人是结直肠癌发病的高危人群。中医药治疗在老年大肠癌的临床治疗中被广泛采用,但对其作疗效评价的较大样本临床对照研究很少。目的:探讨中医药辨证治疗对中老年大肠癌根治术后Ⅱ、Ⅲ期患者生存期的影响。设计、场所、受试者和干预措施:采用同期对照研究方法,收集来自上海

市某医院肿瘤一科、肛肠外科的45 岁及以上老年大肠癌根治术后Ⅱ、Ⅲ期病例,全部病例均行西医常规治疗,以是否自愿接受中医药辨证治疗分为综合治疗组和西医治疗组。主要结果:共有3424 例病例纳入本研究,其中综合治疗组94 例,西医治疗组3330例，综合考虑两组病例的性别、原发部位、病理类型、临床病理分期、化疗周期、放疗以及中药治疗等对预后的影响。得到如下表：

表6.11　中西医治疗大肠癌预后效果比较表

组别	疗效				
	痊愈	显效	好转	无效	合计
治疗组	13	21	51	9	94
对照组	30	670	1870	760	3330

研讨问题

1. 根据数据分析表，如果直接看中医治疗的疗效比例和西医治疗的疗效比例，你会有怎样的看法？设计假设检验，给出统计量，完成统计分析，得到结论。

2. 如果考虑到疗效是用单项有序数据表示的，使用Ridit分析会得到怎样的分析结果？

3. 中医比较关注患者体质症状情况，比如常见的Karnofsky评分，对病人术后每月生活质量调查表随访疾病影响情况进行评分,涉及到的影响侧面有身体和心理社会学,焦郁症,体重减轻情况,包括腹泻、进食、休息、工作能力和睡眠等生活质量等问题。如果和关键临床测量变量血清碱性磷酸酶水平进行比较,请根据以上疗效的不同级别所对应的生存时间，研究Karnofsky 评分和临床测量变量对预后生存时间的影响，特别是对生存时间在前30%的病人的影响，综合分析中西医治疗对病人恢复的整体影响。