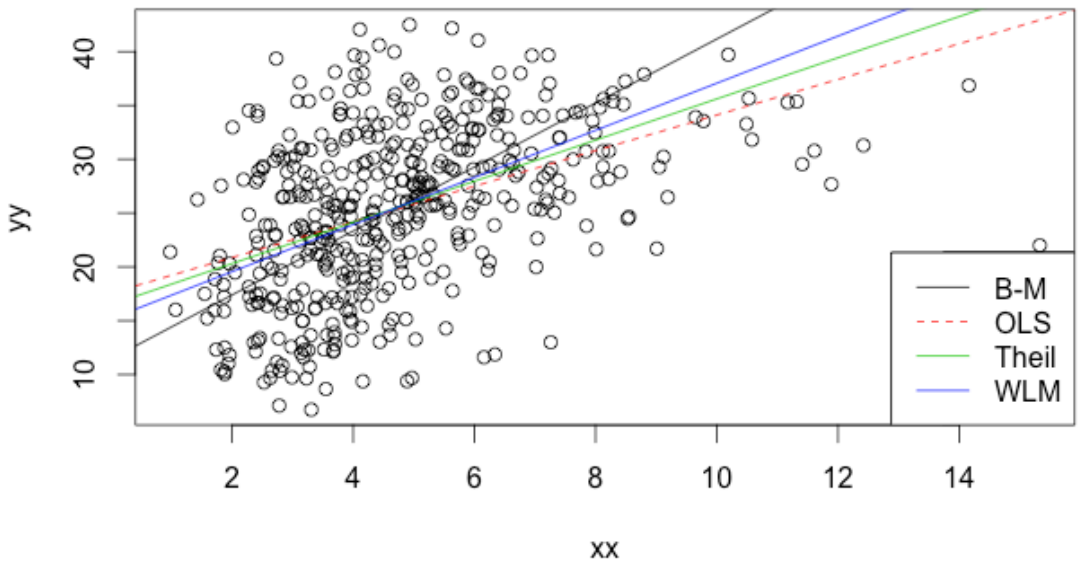
非参数统计第 13 次作业

王健桥 2013202552

数据文件为 **saheart.txt**

分别用 Brown-Mood, OLS, Theil 以及 WLS 做出的回归图像如下：



用自己编写的函数将 OLS，Theil 以及 WLS 分别进行 Brown-Mood 检验，得到结果如下：

$$\mathrm{H}_0 : \alpha = \alpha_0, \beta = \beta_0 \leftrightarrow H_1 : \alpha \neq \alpha_0 或 \beta \neq \beta_0$$

OLS

BM_Rfun(cyx[1],cyx[2],xx,yy)

统计量为 6.796537

[1] "P 值为"

[1] 0.03343111

THEIL

BM_Rfun(intercept,slope,xx,yy)

3.160173

[1] "P 值为"

[1] 0.2059573

WLS

> BM_Rfun(coef(WLM)[1],coef(WLM)[2],xx,yy)

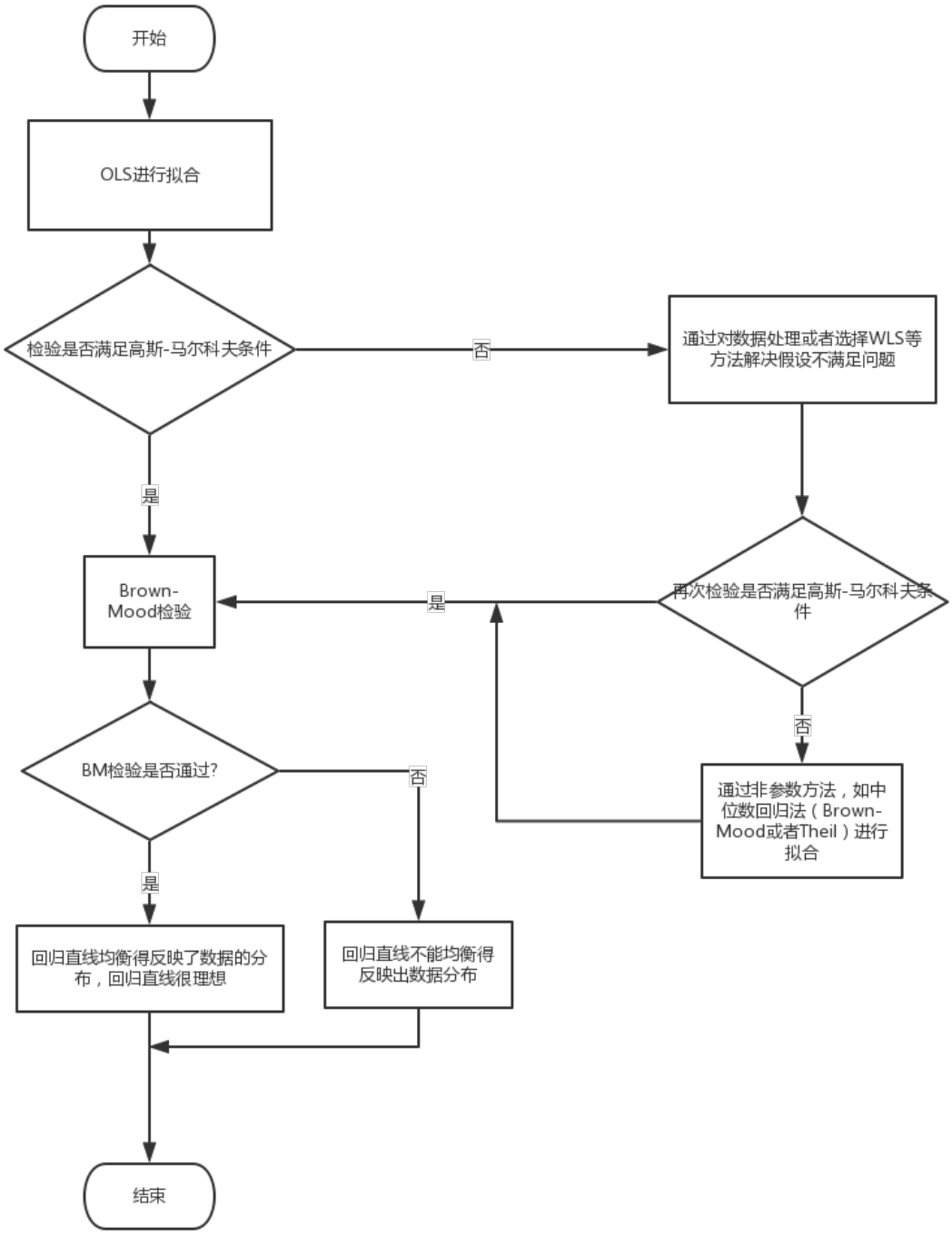
[1] 0.5627706

[1] "P 值为"

[1] 0.7547375

可知 OLS 没有通过 BM 检验，而 THEIL 和 WLS 通过了检验

流程图：



6.3

对于裁判判决的一致性检验，我们采用多变量 Kendall 协和系数检验的方法。

H_0：评委的打分不相关

H_1：评委的打分是相关的，即具有一致性

输入全部的数据，进行检验

Kendall's coefficient of concordance Wt
Chisq(9) = 33.2
 p-value = 0.000123

p 值很小，即能拒绝原假设。

考虑到数据的现实背景，我们更加关注比赛中对排名前列的选手打分情况，所以选取总分在中位数以上的选手所得得分进行一致性检验：

```
A1<-colSums(ratings)
A_num <- which(A1>median(A1))
ratings2<-ratings[,A_num]
kendall(t(ratings2), correct = TRUE)
```

得到结果，
Kendall's coefficient of concordance Wt
Chisq(4) = 20.2
 p-value = 0.000451

依旧能拒绝原假设，则我们得出结论，各个评委之间的判断具有一致性。

6.4

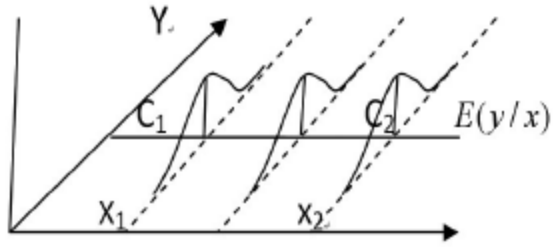
进行 Kappa 一致性检验，原假设为两种方法不一致

Z = 4.3019, p-value = 8.466e-06
95 percent confidence interval:
 0.2430741 0.5918773
sample estimates: 0.4174757

"Moderate agreement"
P 值很小，即可以拒绝原假设，说明两位医生的治疗方法具有一致性。统计量表明是具有中等一致性。

6.8

（1）分位数回归是对普通最小二乘的一种扩充，对于假设因变量 Y 和自变量 X 之间存在着相关关系，由于 Y 是随机变量，对于 X 的各个确定值 x，Y 都有对应的分布 F(y/x)。



直接确定 F(y/x)是困难的，作为一种近似，普通最小二乘回归方法转而确定 E(y/x)。

而分位回归是建立因变量 Y 与自变量 X 的条件分位模型，即

$$Q_Y(\tau \mid X) = f(X)$$

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，中位数拟合就是 0.5 分位回归。

（2）对于分位回归的最优化问题表示形式为：

$$\arg\min_{\beta \in R^n} \left[\sum_{i \in \{i: y_i \geq f(x)\}} \tau \mid y_i - f(x_i) \mid + \sum_{i \in \{i: y_i \leq f(x)\}} (1-\tau) \mid y_i - f(x_i) \mid \right]$$

对于如上的表示形式，我们可以看到分位数回归用的是最小化残差的绝对值形式，这一点与绝对值形式的性质有关，为了更好的说明这一点，不妨设我们分位数为 0.5，即此时为中位数回归，同时有一组样本 $y_1, \ldots, y_n$ ，对于一组样本，$\min \sum (y_i - \xi)^2$ 中，均值是最小值解，而中位数则是 $\sum \mid y_i - \xi \mid$ 的最小值解同样，可以证明，

$\left[\sum_{i \in \{i: y_i \geq f(x)\}} \tau \mid y_i - \xi \mid + \sum_{i \in \{i: y_i \leq f(x)\}} (1-\tau) \mid y_i - \xi \mid \right]$的取最小值时，解为$\tau$ 分位数，所以要寻找分位数的回归，我们采取了最小化绝对值的方式。

当 $f(x_i) = x_i^T \beta$ 时，即为分位数的线性回归。由于绝对值形式的目标函数无法得到解析解，因此只能通过数值方法反复迭，得到最优解。线性规划可以用单纯形法，而非线性规划可以用可以用其他方法解决

（3）(1)当存在显著的异方差等情况最小二乘法估计稳健性非常差。分位回归对模型中的随机误差项不需做具体的分布假定，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和自变量间的关系，因此分位回归体现了数据驱动的建模思想

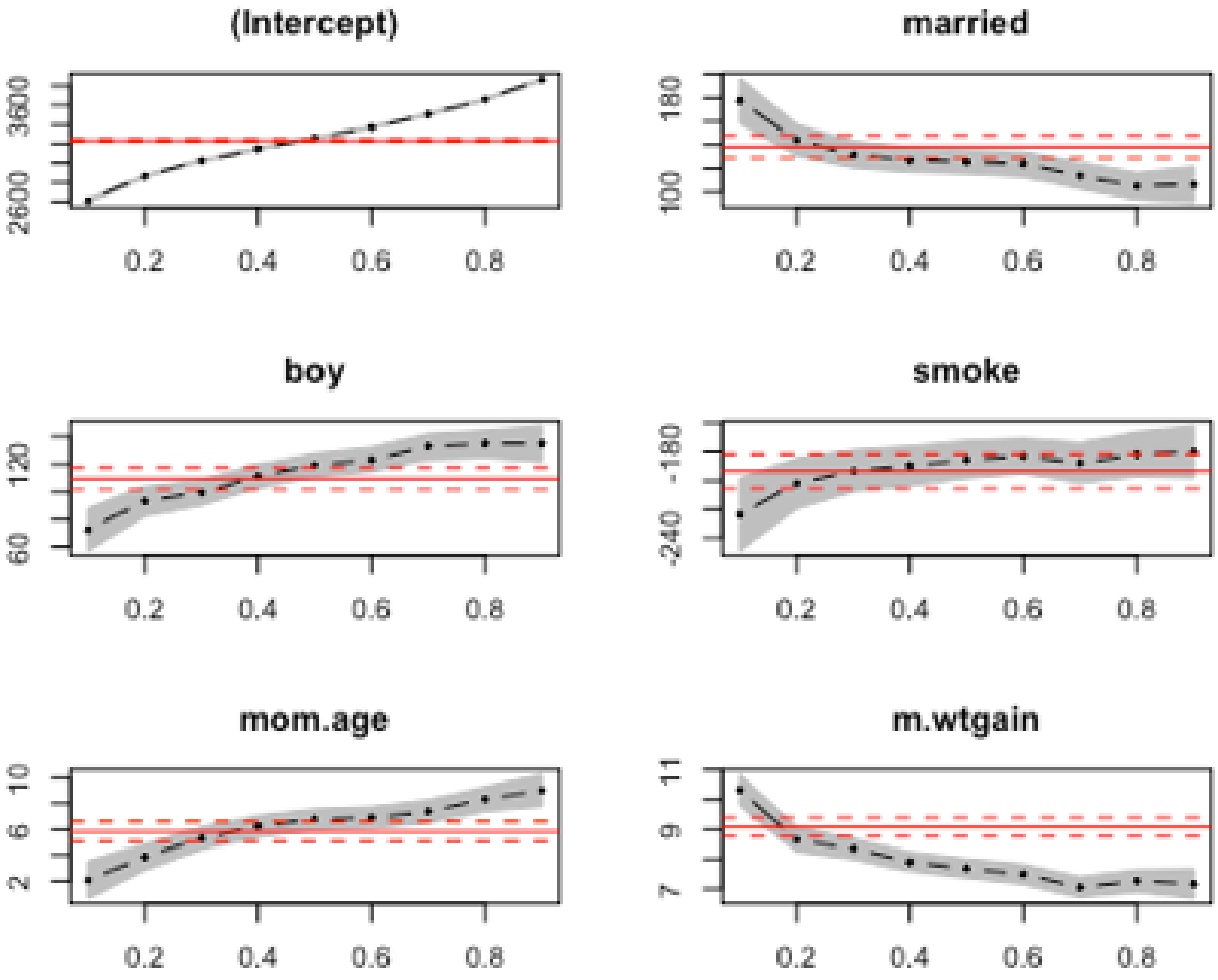
（3）分位回归对分位数τ 进行回归，于是对异常值不敏感，模型结果比较稳定；而 OLS 是对均值进行回归，易受异常值影响。

（4）分位回归可解出不同分位数模型，能更全面的体现分布特点。而 OLS 只能表示均值的变化。

（4）

```
attach(infant_weight) fit1<-
rq(weight~married+boy+smoke+mom.age+m.wtgain,method="pfn",
tau=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)) summary(fit1)
```

得到以下结果



：图中红色实直线及其上下对应虚直线为普通最小回归所得回归系数及对应置信水平 0.05 的置信区间，分位数越大，表示出生时的体重越大，分位数越小，出生时体重越小。

- 图一(intercept)可以解释为一个孕前未婚、不吸烟、年龄为 27 岁、孕期体重增加 30 磅的母亲生下的女孩体重的分位数估计。
- 图 2（married）说明母亲产前婚姻状况对初生婴儿体重有较大正相关，且对于初生婴儿体重较轻的形成的这种关系非常明显；
- 图 3（boy）说明男性初生婴儿体重显著高于女性初生婴儿，但初生婴儿体重较轻时（分位数较小时）这种差距较初生婴儿体重较重时要小；
- 图 4（smoke）说明母亲产前吸烟状况对初生婴儿体重有较大负相关，且对初生婴儿体重较轻的形成影响更加明显；
- 图 5（age）说明母亲孕前年龄对初生婴儿体重有较大正相关，即母亲年龄越大（47 岁为图中上限）初生婴儿体重越大，且母亲产前年龄越小越易生产低体重儿童；
- 图 6（wtgain）说明母亲孕期体重增加值与初生婴儿体重有较大正相关，且对于初生婴儿体重较轻的形成影响非常明显，当分位数越高，即出生体重越重时，这种影响越弱。

参考文献及代码附录

分位数回归的理论再说明及实例分析，乔舰，李再兴

```
data1<-read.table(file = "/Users/w/Desktop/junior_STAT_FIRST_SEM/nmSTAT/数据盘/各章数据/第 6 章/saheart.txt", header = TRUE)

#####普通回归

attach(data1)

fit1<-lm(adiposity~ldl)

#plot(fit1)

############Brown-Mood 方法的中位数回归

cyx = coef(fit1)

yy = adiposity

xx = ldl

y_median = median(yy)

x_median = median(xx)


xx1 = xx[xx>median(xx)]

xx2 = xx[xx<median(xx)]


yy1 = yy[xx>median(xx)]##

yy2 = yy[xx<median(xx)]##


y1_median = median(yy1)

y2_median = median(yy2)


x1_median = median(xx1)
```

```
x2_median = median(xx2)


beta = (y1_median - y2_median)/(x1_median - x2_median)

alpha = median(yy - beta*xx)###### 这个地方注意

############################追求拟合

#########Theil 进行中位数估计

###################如果有节

l2<-NULL

if(length(data1$ldl) != length(unique(data1$ldl))){

   l2 = NULL

   v1 = unique(data1$ldl)

   for( i in 1:length(v1))

    l2<-c(l2,median(data1$adiposity[data1$ldl ==v1[i]]))

   n <- length(l2)

}

##################

combos <- combn(n, 2)

i.s <- combos[1,]

j.s <- combos[2,]

Y.num <- l2[j.s] - l2[i.s]

X.dom <- v1[j.s] - v1[i.s]

Z.p <-  qnorm(alpha/2, lower.tail=F)

N <- (n*(n-1))/2

s <- (Y.num/X.dom)
```

```
C.stat <- sum(sign(s))

slope <- median(s, na.rm=TRUE)

intercept <- median((l2[ ] - slope*v1[ ]), na.rm=TRUE)

############WLS

WLM<-lm(adiposity~ldl,weights = 1/(ldl)^2)########这个加权最小二乘有待考证

plot(xx,yy)

abline(alpha, beta,col = 1)#####

abline(c(cyx),lty = 2,col = 2)

abline(intercept,slope,col =3)

abline(coef(WLM), col = 4)

legend("bottomright",c("B-M","OLS","Theil","WLM"),lty = c(1,2,1,1),col = 1:4)

##############################

BM_Rfun = function(intercept,slope,x,y){

  X_median = median(x)

  l1 = NULL

  l1 <-c((x[]<X_median & y[] > intercept + slope*x[]))

  l2<-c((x[]>X_median & y[] < intercept + slope*x[]))

  T_STATISTICS <-((sum(l1)-length(l1)/4)^2+(sum(l2)-length(l2)/4)^2)*8/length(l1)

  pvalue = pchisq(T_STATISTICS,2,lower.tail = FALSE)

  result<-list(t=T_STATISTICS,info=paste("P 值为"),pvalue)

  result

}

BM_Rfun(cyx[1],cyx[2],xx,yy)#OLS

BM_Rfun(intercept,slope,xx,yy)#Theil

BM_Rfun(coef(WLM)[1],coef(WLM)[2],xx,yy)#WLS
```

```
detach(data1)

################################

install.packages("irr")

library(irr)

ratings<-read.csv(file = "/Users/w/Desktop/junior_STAT_FIRST_SEM/nmSTAT/12-
data.csv",header = T)

ratings<-ratings[!is.na(ratings)]

ratings = matrix(ratings,nrow = 12, ncol =10)

kendall(t(ratings), correct = TRUE)

#####################################

A1<-colSums(ratings)

A_num <- which(A1>median(A1))

ratings2<-ratings[,A_num]

kendall(t(ratings2), correct = TRUE)

#ratings.rank <- apply(ratings, 2, rank)

########################6.4####

D_Matrix<-matrix(c(40,5,25,30),2,2,byrow = T)

library(fmsb)

Kappa.test(D_Matrix)

############ p 值很小，具有显著的一致性

############6.8########

install.packages("quantreg")

library(quantreg)

library(SparseM)

infant_weight<-read.table(file = "/Users/w/Desktop/junior_STAT_FIRST_SEM/nmSTAT/数据盘/
各章数据/第 6 章/infant-birthweight.txt",header = T)
```

```
attach(infant_weight)

plot(m.wtgain,weight,xlab = "m.wtgain",ylable = "weight",type = "n")

points(m.wtgain,weight)

taus = seq(0.1,0.9,0.1)

f = coef(rq(weight~m.wtgain,tau = taus));

for (i in 1:length(taus)){

  abline(f[,i][1],f[,i][2],lty = 2,color = i)

}

abline(lm(weight~m.wtgain),lty = 9)

legend(3000,700,c("mean","median","otherquantile"),lty = c(9,1,2))

fit1<-rq(weight~married+boy+smoke+mom.age+m.wtgain,method="pfn",
tau=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9))

summary(fit1)

plot(summary(fit1, alpha=0.05))

detach(infant_weight)
```