**生存分析期末报告**

**曾子轩**

**17344011**

**目录**

[**1、问题背景 3**](#_Toc45810100)

[**2、数据处理 3**](#_Toc45810101)

[**2.1数据读取 3**](#_Toc45810102)

[**2.2缺失值处理 4**](#_Toc45810103)

[**3、问题探索 4**](#_Toc45810104)

[**3.1冠状动脉心脏病的病因研究 5**](#_Toc45810105)

[**3.1.1病因筛选 5**](#_Toc45810106)

[**3.1.2 Cox模型 6**](#_Toc45810107)

[**3.1.3 AFT模型 9**](#_Toc45810108)

[**3.2心血管疾病致死研究 10**](#_Toc45810109)

[**3.2.1数据处理及扩展 10**](#_Toc45810110)

[**3.2.2 Extend 模型 11**](#_Toc45810111)

[**4、结论 12**](#_Toc45810112)

[**4.1 CHD病因分析 12**](#_Toc45810113)

[**4.2 死亡风险分析 12**](#_Toc45810114)

[**5、附录 13**](#_Toc45810115)

[**5.1数据及数据说明来源 13**](#_Toc45810116)

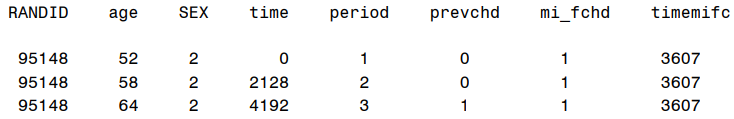
[**5.2 R代码 13**](#_Toc45810117)

**1、问题背景**

Framingham Heart Study是一项对美国马萨诸塞州弗拉明翰社区的志愿者们进行的心血管疾病病因长期性研究。这项研究始于1948年，每两年对5000名左右的参与者进行一次检查，并持续跟踪所有参与者的心血管健康状况。检查数据包括血压、血液化验、肺功能、吸烟史、心电图等。

本报告使用Framingham Heart Study的部分数据集进行生存分析问题的探索。该数据集在12年的时间跨度上，分三个检查阶段收集了参与者的临床检查数据，也记录了心绞痛、心肌梗死、动脉血栓形成性梗死、脑溢血、死亡等若干种事件。由于有多个检查阶段，每个参与者均有1至3个观察结果。

下面以一些示范数据来说明部分字段的含义。



可以看到该参与者的编号为95148（RANDID），第一行数据显示第一阶段（period）时间为0（time），即研究刚开始时该参与者年龄为52（age），性别为女（SEX），本阶段未患上CHD（prevchd），但随着跟踪研究的推进最终患上（mi\_fchd），患上时间为3607（timemifc）。其它阶段或事件可以依此类推。

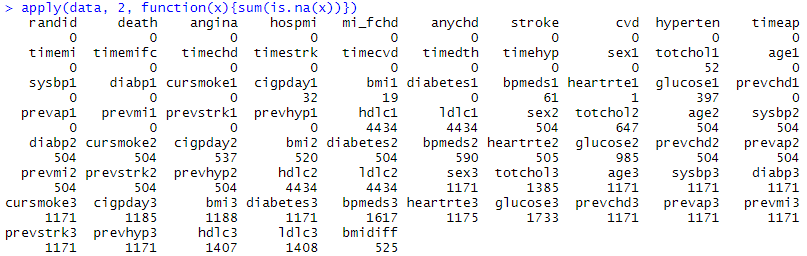
**2、数据处理**

**2.1数据读取**

首先导入数据集。这个数据集包含了4434个样本（行），即记录了4434个参与者，每个参与者有75个字段变量（列）。可以看出与示范数据不同的是，每一个参与者只有一行数据，不同检查阶段的数据将以后缀“1、2、3”来区分，如第一阶段的年龄将标识为“age1”，第二阶段的性别标识为“sex2”。

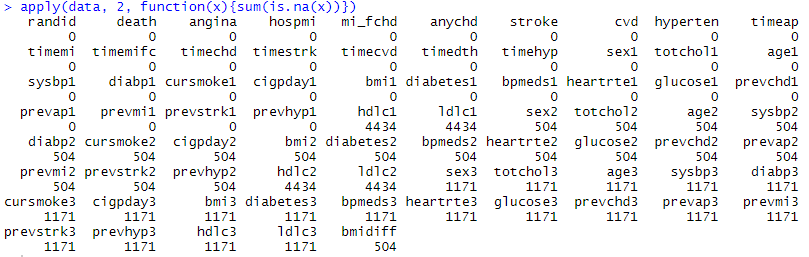
**2.2缺失值处理**

接下来查看变量的缺失情况：



对于空缺的变量，可以分为两种情况：正常缺失（参与者已死亡或者未参加该阶段），这种情况下由于对应阶段的所有字段均空缺，我们将不予处理，模型将自动剔除这些缺失值；异常缺失，即非正常的缺失，此时考虑使用中位数（数值变量）或众数（分类变量）来填补。

这里我们采取的做法是：先通过age变量（根据材料可以知道此变量没有异常缺失）筛选出异常的缺失，然后用中位数或众数填补这些缺失值。注意hdlc和ldlc是例外，根据材料可知它们只适用于第三检查阶段。以下是处理后的数据缺失情况，可以看到现在只剩下正常缺失（各阶段字段的缺失均与age变量的情况保持一致）：



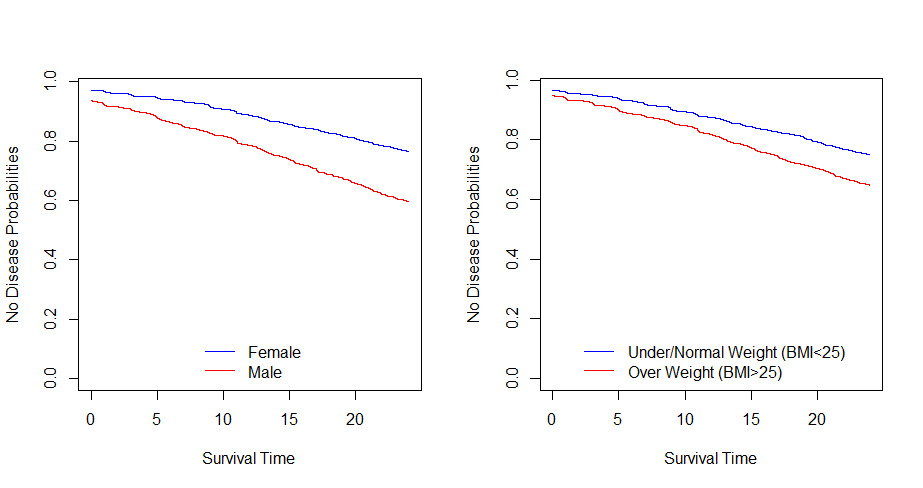
**3、问题探索**

**3.1冠状动脉心脏病的病因研究**

我们对冠状动脉心脏病（CHD）的成因感兴趣，下面将试图从数据给出的变量中找出病因，并使用多种生存模型来概括这些病因与CHD患病风险之间的关系。为方便研究，本节所有变量均使用【首次发现CHD阶段的记录】（若始终未发病则使用最后的非空记录）。

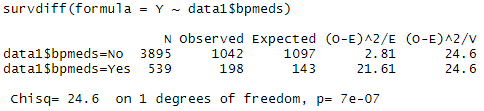
**3.1.1病因筛选**

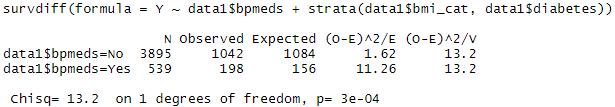
以Sex（性别）、BMI（身体质量指数，需要离散化）两个变量为例，我们使用Kaplan Meier Plot来初步探索病因：



从KM图可以看出，两个变量的取值对患病风险均可能有影响，具体表现为：男性的患病概率大于女性，轻体重/正常体重人群的患病概率大于超重人群。因此在未考虑干扰因子的情况下，可以初步考虑将这两个变量加入模型。

对于某些变量可能存在这样的情况：从分组KM图不能看出患病概率存在明显差别，又或者怀疑有干扰因子。这时可以考虑使用Log Rank Test或者Stratified Log Rank Test。 下面首先使用Log Rank Test来检验Bpmeds（是否使用降压药）对患病风险是否有影响，然后又针对一些可能的干扰因子进行了一次分层的Log-Rank-Test，结果如下：





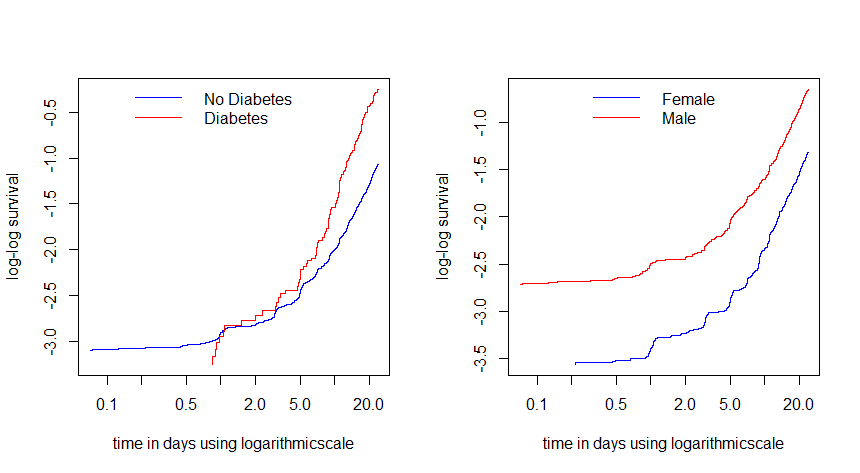
从分层和不分层两次检验结果的p值来看，因子BMI、Diabetes（是否患糖尿病）确实对Bpmeds因子的作用有干扰，但我们仍能得出是否服用降压药对患病风险有显著影响的结论。

**3.1.2 Cox模型**

从3.1节可以得到一些对患病风险有影响的变量，本节将用它们来构建Cox模型。

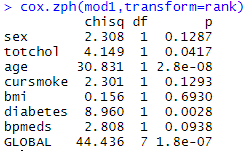
**3.1.2.1 PH假设**

需要注意的是，Cox模型中的变量均必须满足Proportional Hazards Assumption，即PH假设。以Diabetes和Sex变量为例，我们绘制Log-Log Plot来探究变量是否符合PH假设：

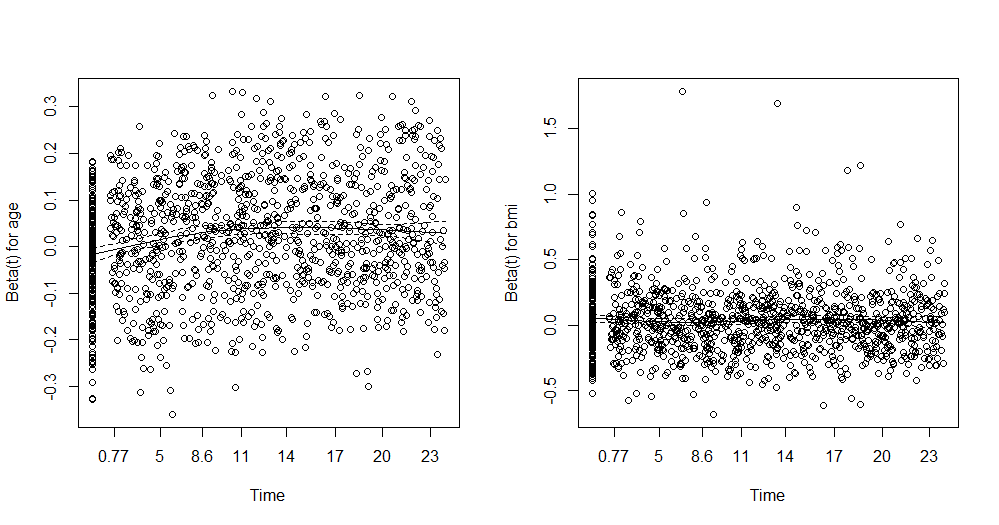


可以看到第一张图曲线有交叉，第二张图曲线则接近平行状态，初步可以判断Diabetes不符合PH假设，Sex符合PH假设。

依据双Log图只能对一些离散变量的PH假设进行粗略的判断，更严谨全面的结果需要通过假设检验来给出。我们首先依据3.1节得到的变量构建Cox模型，然后对PH假设进行检验，结果如下：



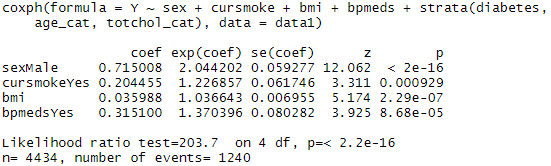
可以看到Sex变量对应的P值为0.1287，Diabetes变量对应的P值为0.0028；设定显著性水平为0.05，我们可以得到与双Log图分析结果完全一致的结论。为了明确本检验的原理，以Age和BMI为例绘制如下的Schoenfeld残差图：



可以看到Age变量对应的图中的点比较分散且拟合曲线不是直线，判断残差与时间相关性较大，PH假设不成立；相反，BMI变量对应的图中的点集中于一水平直线上，说明相关系数接近0，残差与时间相关性较小，因此判断PH假设成立。这与假设检验得到的结论完全一致。

**3.1.2.2分层Cox模型**

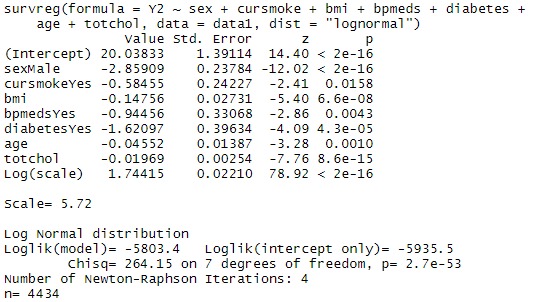
依据上述的结论，我们构建分层的Cox模型，进行分层的变量将是那些不符合PH假设的变量（对于其中的数值变量将进行手动分割）：



观察到p值一列均较小，说明这几个病因对CHD发病的影响都是显著的，这也印证了3.1筛选的有效性。再观察Hazard Ratio，即exp（coef）一列，可以得到病因影响的量化表现，以Sex一行为例，HR值为2.044说明在本模型下，男性的患病风险率约为女性的2.044倍。依此类推，我们可以得到更多结论：吸烟会将风险提高至正常人的1.227倍，一单位的BMI提升会提高3.7%的风险，服用降压药说明有着等同于常人1.370倍的患病风险（由于降压药也是冠心病的治疗药物之一，这里的风险更多代表着一种推断关系，而不是因果关系）。

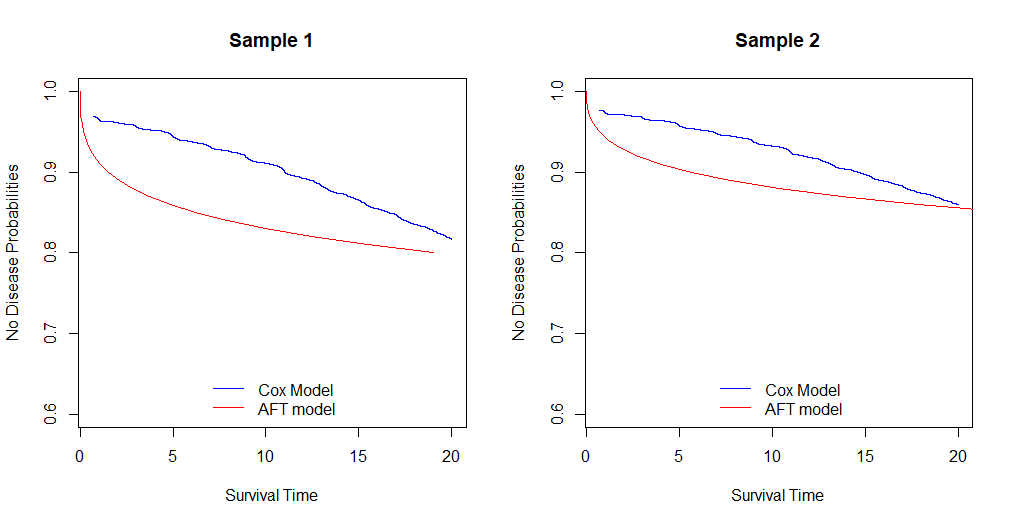
**3.1.3 AFT模型**

接下来我们使用相同的病因来构建加速失效（AFT）模型，以此来量化病因与患病风险的关系。这里使用的是AFT系列模型中的一个代表：Log-Normal模型。



可以看到所有变量系数都是显著的，也就是说所有病因对于患病风险的影响都是显著的，这与Cox模型的结论一致。计算每个系数的exp值可以得到在本模型下，男性的预测生存（不患病）时间为女性的exp（-2.859），即0.057倍；吸烟者的预测生存时间为正常人的0.557倍；提高一单位BMI将导致生存时间变为86.3%；其它变量依次类推。

为了进一步比较Cox模型和AFT模型的结果，我们从数据集中随机挑出两位参与者，分别用不同的模型对他们的生存时间、生存概率进行预测，并绘制出了不同模型下的生存曲线图，结果如下：



可以明显地看到，两种模型的生存曲线趋势完全不同。这是因为加速失效模型的假设是病因的影响将逐渐地（加速地）失去效果，而Cox模型则预设了HR ratio恒定不随时间改变。

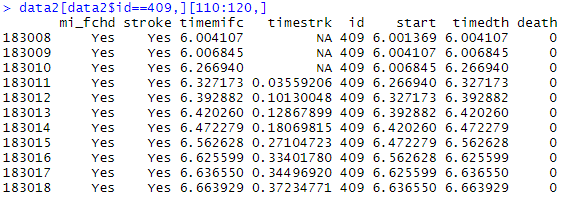
**3.2心血管疾病致死研究**

接下来我们想要研究冠心病（fmi\_chd）和中风（Stroke）的患病时间和死亡风险的关系。

**3.2.1数据处理及扩展**

现在我们来增加【患病后的存活时间】这一指标。显然，这一指标是依赖于时间的，因此需要对数据进行处理以及扩展。具体过程是：首先我们只保留至少患两种疾病之一的参与者的相关数据，然后针对每一个死亡时间，为每一位参与者创建一个数据，然后用对应的死亡时间减去发病时间，即得到患病后的存活时间。存活时间应当为非负，因此我们将负的时间跨度重设为缺失值。

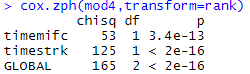
借助R语言中的survSplit函数，我们可以完成以上数据扩展的操作。选取一位参与者，打印其部分数据如下：

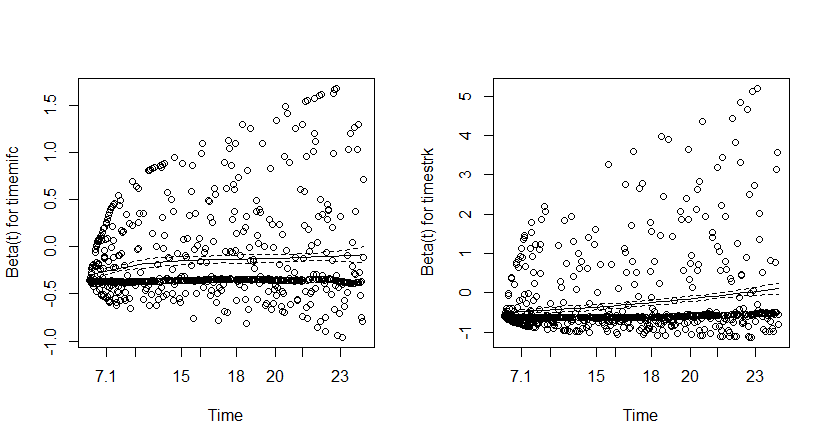


可以看到，Timemifc和Timestrk分别代表了患两种病后的存活时间。

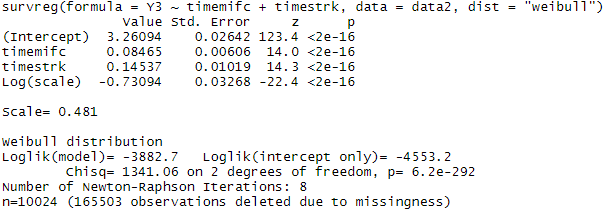
**3.2.2 Extend 模型**

首先对这两个变量做PH假设的检验，并绘制Schoenfeld残差图：





可以明显看出，它们均不符合PH假设。因此Cox模型是不合适的，我们将使用AFT模型。这次我们使用的是Weibull模型：



可以看到，模型的各系数均是显著的。从模型可以得到：患冠心病的时间每增加一单位，预测生存时间将变为exp(0.08465)，即1.088倍；患中风的时间每增加一单位，预测生存时间将变为exp(0.14537)，即1.156倍。

**4、结论**

**4.1 CHD病因分析**

根据3.1节的模型结果可知：男性、吸烟、高身体质量指数、服用降压药、糖尿病、高龄、血液高胆固醇均会导致CHD患病风险的增加，其中尤以性别影响最大。从量化的角度看，以吸烟习惯为例，Cox模型告诉我们吸烟者的患病风险约为常人的1.2倍，AFT模型则告诉我们吸烟者的推测患病时间点比常人提早约50%。两个模型的差异主要在于对于数据的预设不同，AFT模型假设生存时间取决于基准生存时间和加速因子，Cox模型则假设Hazard Ratio恒定不变，所以它们所呈现出来的生存曲线有着较大差异。

**4.2 死亡风险分析**

根绝3.2节的结果可知：冠心病患者每存活一单位时间，预测生存时间就会增加8.8%，而患中风的存活时间每增加一单位，预测生存时间将增加15.6%。因此，对于冠心病和中风，要特别注意其带来的急性死亡的风险。

注意到此模型存在一些缺陷，如没有排除其它急性疾病对死亡率的影响，这将使结果产生一定的偏移。

**5、附录**

**5.1数据及数据说明来源**

数据来源：handouts\_fhs.csv

数据说明文件：Framingham Longitudinal Data Documentation\_v2.pdf

**5.2 R代码**

（代码运行环境为R version 4.0.2）

## 读取数据

data<-read.csv("handouts\_fhs.csv")

library(survival)

## 处理缺失值

# 查看数据缺失情况

data[data=='']<-NA

apply(data, 2, function(x){sum(is.na(x))})

# 用中位数或众数填补异常的缺失值

getmode<-function(v)

{uniqv<-unique(v);uniqv[which.max(tabulate(match(v, uniqv)))]}

fillna<-function(v, method='median', period=1)

{if(method=='mode') {x<-getmode(v)} else {x<-median(v, na.rm = TRUE)}

if(period==2) {v[is.na(v)&!(is.na(data['age2']))]<-x}

else if(period==3) {v[is.na(v)&!(is.na(data['age3']))]<-x}

else {v[is.na(v)&!(is.na(data['age1']))]<-x}

v}

for (col in list('totchol1','cigpday1','bmi1','bpmeds1','heartrte1','glucose1'))

{v<-as.matrix(data[col])

if(col=='bpmeds1') {data[col]<-fillna(v,method='mode')}

else {data[col]<-fillna(v)}}

for (col in list('totchol2','cigpday2','bmi2','bpmeds2','heartrte2','glucose2','bmidiff'))

{v<-as.matrix(data[col])

if(col=='bpmeds2') {data[col]<-fillna(v,method='mode',period=2)}

else {data[col]<-fillna(v,period=2)}}

for (col in list('totchol3','cigpday3','bmi3','bpmeds3','heartrte3','glucose3','hdlc3','ldlc3'))

{v<-as.matrix(data[col])

if(col=='bpmeds3') {data[col]<-fillna(v,method='mode',period=3)}

else {data[col]<-fillna(v,period=3)}}

# 查看处理后的数据缺失情况

apply(data, 2, function(x){sum(is.na(x))})

## CHD病因研究

# 构造数据集

data1<-data.frame(matrix(NA,4434,21))

colnames(data1)<-c('anychd','timechd','sex','totchol','age','sysbp','diabp','cursmoke','cigpday','bmi','diabetes',

'bpmeds','heartrte','glucose','prevchd','prevap','prevmi','prevstrk','prevhyp','hdlc','ldlc')

rownames(data1)<-rownames(data)

data1$anychd<-data$anychd;data1$timechd<-data$timechd

row3<-!is.na(data$prevchd3)&data$prevchd3=='Yes'

row2<-!is.na(data$prevchd2)&data$prevchd2=='Yes'

row1<-data$prevchd1=='Yes'

row4<-!(row3|row2|row1)

data1[row3,3:21]<-data[row3,56:74]

data1[row2,3:21]<-data[row2,37:55]

data1[row1,3:21]<-data[row1,18:36]

data1[row4,3:21]<-data[row4,18:36]

data1[!is.na(data$prevchd2)&row4,3:21]<-data[!is.na(data$prevchd2)&row4,37:55]

data1[!is.na(data$prevchd3)&row4,3:21]<-data[!is.na(data$prevchd3)&row4,56:74]

# KM Plot

Y<-Surv(data1$timechd, data1$anychd=='Yes')

data1$bmi\_cat<-data$bmi

data1$bmi\_cat[data1$bmi<25]<-'Under/Normal Weight'

data1$bmi\_cat[data1$bmi>=25]<-'Over Weight'

par(mfrow=c(1,2))

kmfit<-survfit(Y~data1$sex)

plot(kmfit,col=c('blue','red'),xlab="Survival Time",ylab="No Disease Probabilities")

legend('bottom', c("Female","Male"),

lty="solid", col=c('blue','red'), bty='n')

kmfit<-survfit(Y~data1$bmi\_cat)

plot(kmfit,col=c('red','blue'),xlab="Survival Time",ylab="No Disease Probabilities")

legend('bottom', c("Under/Normal Weight (BMI<25)","Over Weight (BMI>25)"),

lty="solid", col=c('blue','red'), bty='n')

# Log Rank Test

survdiff(Y~data1$bpmeds)

survdiff(Y~data1$bpmeds+strata(data1$bmi\_cat,data1$diabetes))

# Log-Log Plot

par(mfrow=c(1,2))

kmfit<-survfit(Y~data1$diabetes)

plot(kmfit, fun='cloglog',col=c('blue','red'),

xlab="time in days using logarithmicscale",ylab="log-log survival")

legend('top',c("No Diabetes","Diabetes"),lty="solid",col=c('blue','red'),bty='n')

kmfit<-survfit(Y~data1$sex)

plot(kmfit, fun='cloglog',col=c('blue','red'),

xlab="time in days using logarithmicscale",ylab="log-log survival")

legend('top',c("Female","Male"),lty="solid",col=c('blue','red'),bty='n')

# Test of PH Assumption

mod1<-coxph(Y~sex+totchol+age+cursmoke+bmi+diabetes+bpmeds, data=data1)

cox.zph(mod1,transform=rank)

# Schoenfeld残差图

par(mfrow=c(1,2))

plot(cox.zph(mod1,transform=rank),var='age')

plot(cox.zph(mod1,transform=rank),var='bmi')

# Stratified Cox Model

data1$age\_cat<-cut(data1$age,breaks=c(-Inf,50,60,70,Inf))

data1$totchol\_cat<-cut(data1$totchol,breaks=c(-Inf,200,330,Inf),labels=c("lowq","normalq","highq"))

mod2<-coxph(Y~sex+cursmoke+bmi+bpmeds+strata(diabetes,age\_cat,totchol\_cat), data=data1)

# AFT Model

data1$timechd2<-data1$timechd

data1$timechd2[data1$timechd2==0]<-0.0001

Y2<-Surv(data1$timechd2,data1$anychd=='Yes')

mod3<-survreg(Y2~sex+cursmoke+bmi+bpmeds+diabetes+age+totchol, data=data1, dist='lognormal')

summary(mod3)

# Survival Curve

pattern1<-data1[2000,];pattern2<-data1[3000,]

i<-0;par(mfrow=c(1,2))

for (pattern in list(pattern1,pattern2))

{

i<-i+1

plot(survfit(mod1, newdata=pattern),conf.int=F,

xlim=c(0.7,20),ylim=c(1,0.6),col='blue',

xlab="Survival Time",ylab="No Disease Probabilities")

pct=0:20/100

days=predict(mod3,newdata=pattern,type='quantile', p=pct)

survival=1-pct

lines(days,survival,col='red')

legend('bottom',c("Cox Model","AFT model"),col=c('blue','red'),lty='solid',bty='n')

title(paste("Sample",as.character(i)))

}

## 死亡原因研究

# survSplit

data2<-subset(data,select=c('death','mi\_fchd','stroke','timemifc','timestrk','timedth'))

data2<-data2[data2$mi\_fchd=='Yes'|data2$stroke=='Yes',]

data2$death[data2$death=='Yes']<-1;data2$death[data2$death=='No']<-0

data2$death<-as.numeric(data2$death)

data2<-survSplit(data=data2,cut=data2$timedth[data2$death==1],

end='timedth', event='death', start='start', id='id')

data2$timemifc<-data2$timedth-data2$timemifc

data2$timemifc[data2$timemifc<0]<-NA

data2$timestrk<-data2$timedth-data2$timestrk

data2$timestrk[data2$timestrk<0]<-NA

data2<-data2[!is.na(data2$timemifc)|!is.na(data2$timestrk),]

data2[data2$id==409,][110:120,]

# Extend Model

Y3<-Surv(data2$timedth,data2$death)

mod4<-coxph(Y3~timemifc+timestrk,data=data2)

cox.zph(mod4,transform=rank)

par(mfrow=c(1,2))

plot(cox.zph(mod4,transform=rank),var='timemifc')

plot(cox.zph(mod4,transform=rank),var='timestrk')

mod5<-survreg(Y3~timemifc+timestrk,data=data2,dist='weibull')

summary(mod5)