**1、**

**Rcode：**

## 读取数据

data<-read.csv("wcgsdata.csv")

library(survival)

Y<-Surv(data$Time169, data$Chd69==1)

## 处理BMI,Ncigs0,Chol0,Dibpat0数据

data$Weight0<-data$Weight0 \* 0.45359

data$Height0<-data$Height0 \* 0.0254

data$BMI<-data$Weight0 / (data$Height0)^2

data$Ncigs0[data$Ncigs0!=0]<-"smoker"

data$Ncigs0[data$Ncigs0==0]<-"not smoker"

med <- median(as.numeric(data$Chol0[data$Chol0!='.']))

data$Chol0[data$Chol0=='.']<-med

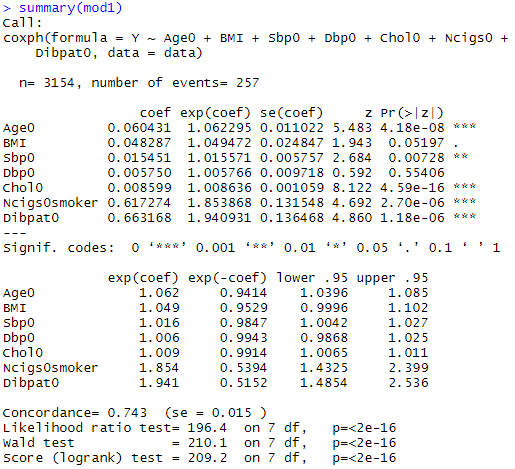
data$Chol0<-as.numeric(data$Chol0)

## Cox Model

mod1<-coxph(Y ~ Age0 + BMI + Sbp0 + Dbp0 + Chol0 + Ncigs0 + Dibpat0, data=data)

summary(mod1)

**输出：**



**结论：**

1）从Pr（>|z|）一列可以有95%的把握推断Age0，BMI，Sbp0，Chol0，Ncigs0，Dibpat0这几个变量对患病风险有影响；从exp（coef）一列则可以看Ncigs0取“smoker”或者其他变量增大会导致患病风险率的增加。

2)从Dibpat0变量的exp（coef）一列可以看出Behavior type A相对type B的Harzard Ratio为1.941，对应的95%置信区间为（1.4854，2.536）。这一结论与Assignment 2中得出的Behavior type A会导致发病率升高的结论吻合。

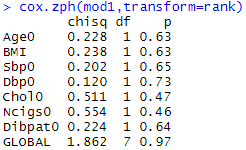
**2、**

**Rcode：**

## Test of PH Assumption

cox.zph(mod1,transform=rank)

**输出：**



**结论：**

可以得出：我们没有95%的把握来否定各变量以及整个模型不符合PH假设。

**3、**

**Rcode：**

## Log-Log Plot

for (col in list("Age0", "BMI", "Sbp0", "Dbp0", "Chol0"))

{

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

med <- median(v, na.rm=T)

v[v<=med]<-1

v[v>med]<-2

kmfit<-survfit(Y ~ v)

plot(kmfit, fun='cloglog',

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

main=c("log-log curves by ",col))

}

for (col in list("Ncigs0", "Dibpat0"))

{

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

kmfit<-survfit(Y ~ v)

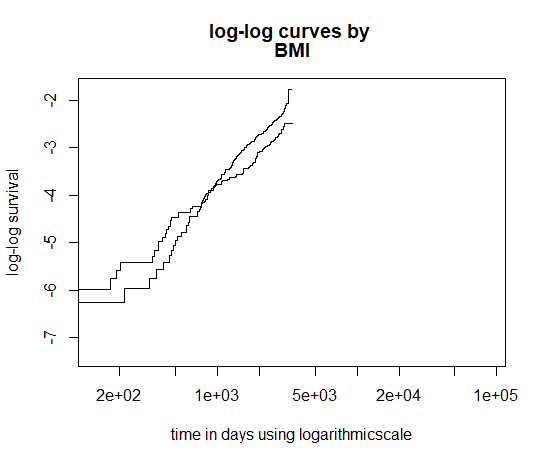
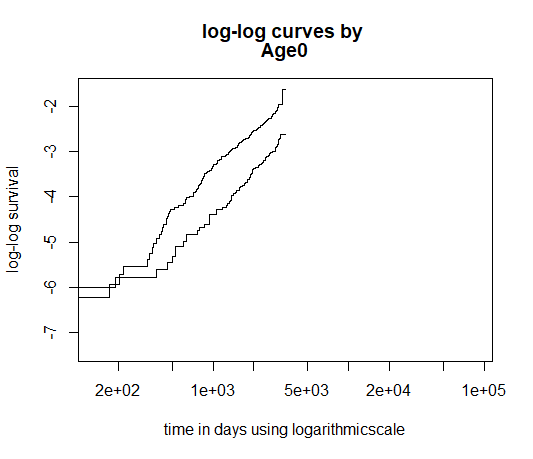
plot(kmfit, fun='cloglog',

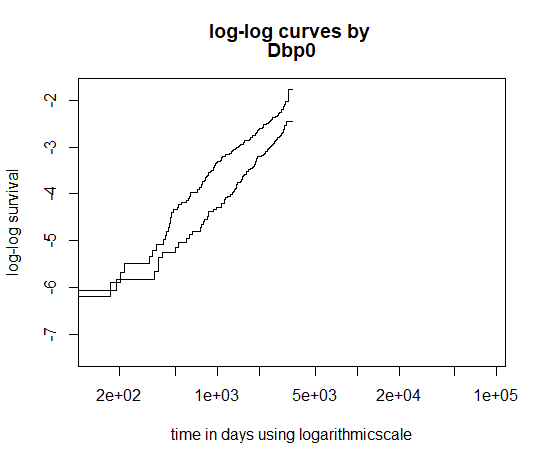
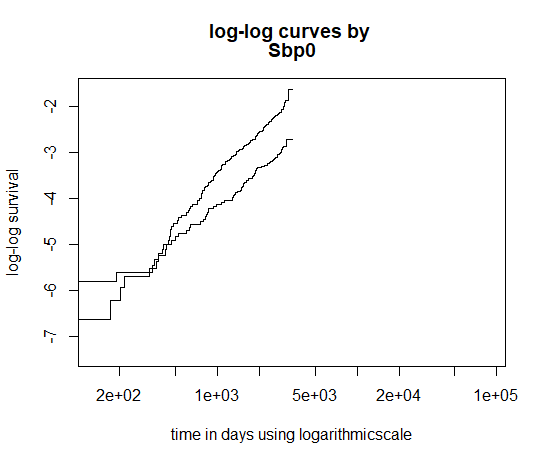
xlab="time in days using logarithmicscale", ylab="log-log survival",

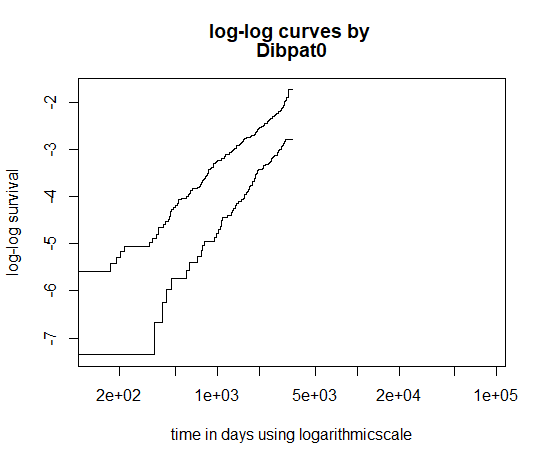
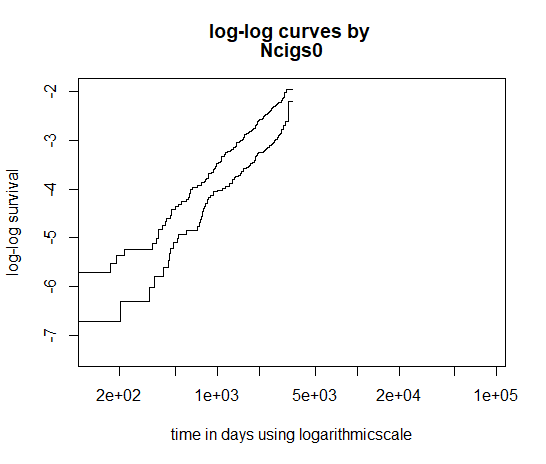
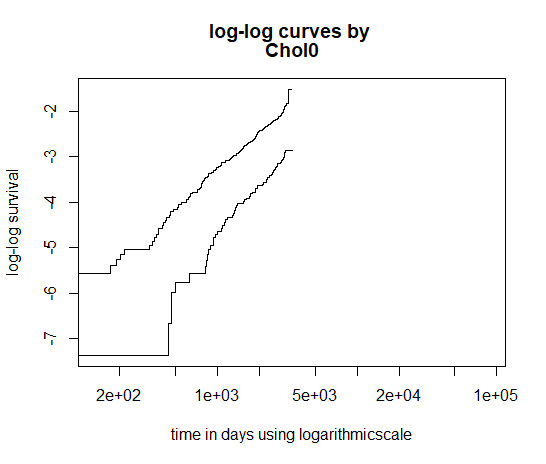
main=c("log-log curves by ",col))

}

**输出：**







**结论：**

由各log-log图可知，Age0，BMI，Sbp0，Dbp0这几个变量并不符合PH假设；Chol0，Ncigs0，Dibpat0则基本符合PH假设。

**4、**

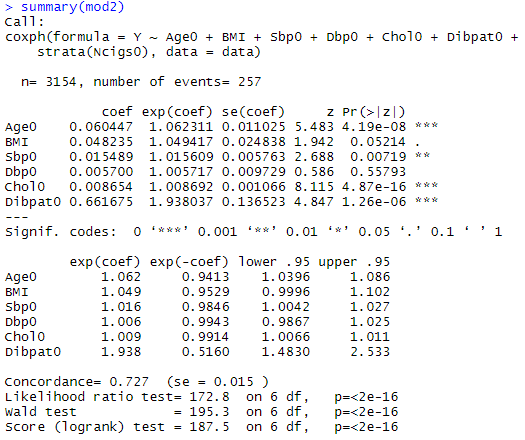
**Rcode：**

## Stratified Cox Model

mod2<-coxph(Y ~ Age0 + BMI + Sbp0 + Dbp0 + Chol0 + Dibpat0 + strata(Ncigs0), data=data)

summary(mod2)

**输出：**



**5、**

**Rcode：**

## Extended Cox Model

data.cp <- survSplit(data,cut=data$Time169[data$Chd69==1],

end='Time169',event='Chd69',start='start',id='id')

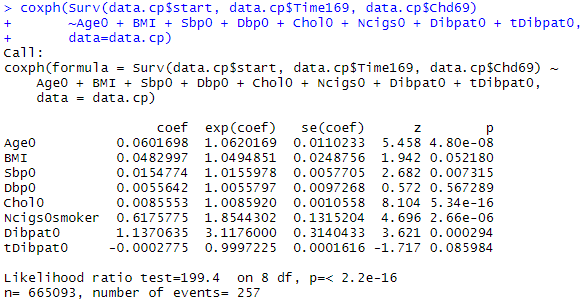
data.cp$tDibpat0 <- data.cp$Dibpat0 \* data.cp$Time169

coxph(Surv(data.cp$start, data.cp$Time169, data.cp$Chd69)

~Age0+BMI+Sbp0+Dbp0+Chol0+Ncigs0+Dibpat0+tDibpat0,

data=data.cp)

**输出：**



**结论：**

tDibpat0变量p值为0.086，说明该系数不显著，没有95%的把握断定Dibpat0×t对风险率有影响。

**6、**

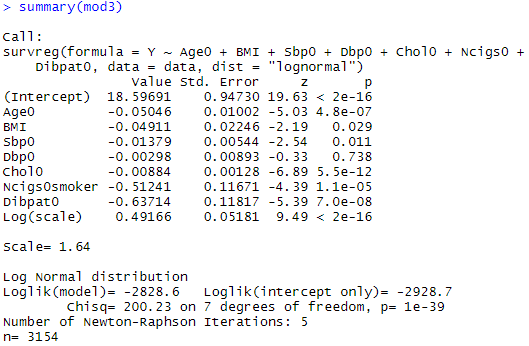
**Rcode：**

## Log Normal Survival Model

mod3<-survreg(Y~Age0+BMI+Sbp0+Dbp0+Chol0+Ncigs0+Dibpat0, data=data, dist='lognormal')

summary(mod3)

**输出：**



**结论：**

从p一列可以有95%的把握推断Age0，BMI，Sbp0，Chol0，Ncigs0，Dibpat0这几个变量对患病率有影响；这一结论与1的Cox Analysis完全相同。

**7、**

**Rcode：**

## Predict

pattern1<-data.frame(Age0=55,BMI=28.5,Sbp0=138,Dbp0=90,Chol0=280,Ncigs0='smoker',Dibpat0=1)

pattern2<-data.frame(Age0=42,BMI=22,Sbp0=120,Dbp0=80,Chol0=180,Ncigs0='not smoker',Dibpat0=1)

summary(survfit(mod1, newdata=pattern1), times=365\*5)

summary(survfit(mod1, newdata=pattern2), times=365\*5)

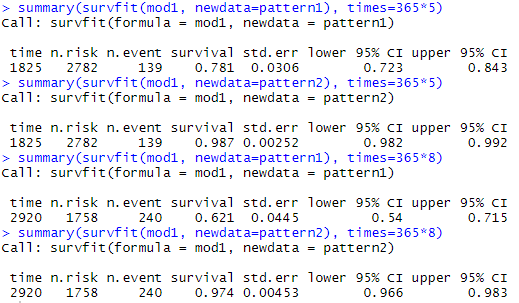
summary(survfit(mod1, newdata=pattern1), times=365\*8)

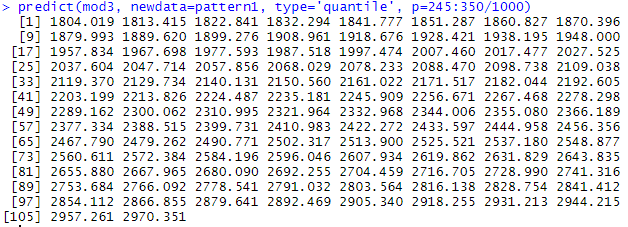
summary(survfit(mod1, newdata=pattern2), times=365\*8)

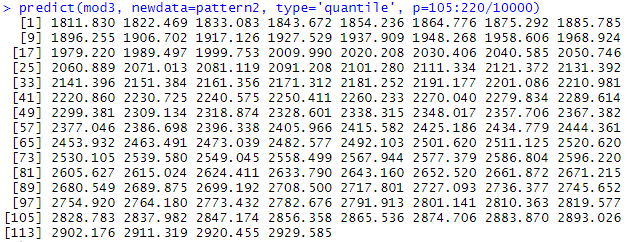
predict(mod3, newdata=pattern1, type='quantile', p=245:350/1000)

predict(mod3, newdata=pattern2, type='quantile', p=105:220/10000)

**输出：**







**结论：**

由前四个输出可以看到，根据Cox模型，A朋友5年（即1825天）内发病的概率为1-0.781，即21.9%，8年（即2920天）内发病的概率为1-0.621，即37.9%；B朋友5年内发病的概率为1-0.987，即1.3%，8年内发病的概率为1-0.974，即2.6%。

由后两个输出可以看到，根据Log-normal模型，A朋友5年内发病的概率大致为247/1000，即24.7%，8年（即2920天）内发病的概率大致为346/1000，即34.6%；B朋友5年内发病的概率大致为106/10000，即1.06%，8年内发病的概率大致为219/10000，即2.19%。