#数据导入以及缺失值处理等

ybdf <- read.table("D:/Document/OneDrive/DailyTask/创新服务大赛/16000.csv",

header = TRUE,sep = ",")

View(ybdf)

str(ybdf)

head(ybdf)

dim(ybdf)

any(is.na(ybdf)) # 检查是否有缺失值

#删除含有缺失值的行

ybdf <- na.omit(ybdf)

dim(ybdf)

#删除相关性较强的变量

for (i in 1:ncol(ybdf)){

if(length(unique(ybdf[, i]))==1){

print(names(ybdf[i]))

}

}

#[1] "个人编码"

#[1] "交易时间YYYY\_NN"

#[1] "住院天数\_SUM"

for (i in 1:ncol(ybdf)-3){

if(length(unique(ybdf[, i]))==1){

ybdf <- ybdf[,-i]

}

}

head(ybdf)

dim(ybdf)

str(ybdf)

cor(ybdf)

#主成分降维

#确定主成分的个数

library(psych)

#碎石图

fa.parallel(ybdf[,-79],fa="pc",n.iter = 100,show.legend = FALSE,

main = "Scree plot with parallel analysis")

图表, 文本

中度可信度描述已自动生成

通过碎石图中特征值大于1准则（水平线）和100次模拟的平行分析（虚线）都表明可以保留18、19、20个主成分，后续可通过累计贡献以及结合模型来确定最终主成分个数。

#确定主成分个数

PCA <- princomp(ybdf[,-79],cor=T)

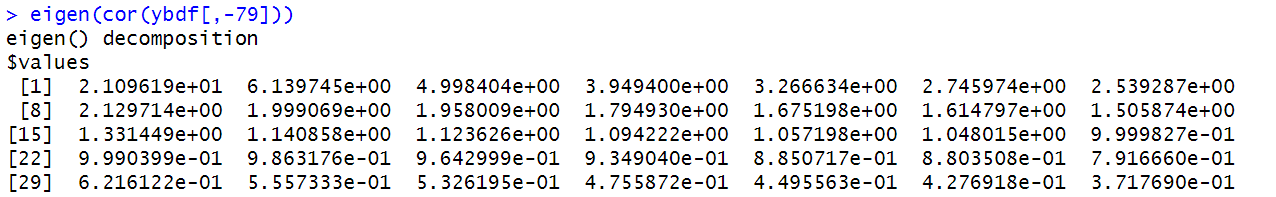
summary(PCA)

#特征根

eigen(cor(ybdf[,-79]))

图片包含 日历

描述已自动生成



可以看到，第一主成分解释了27%的方差，第二主成分解释了7.9%的方差……前20个主成分的特征根>1，同时累计贡献率可达到82.3%，表明前20个主成分基本包含了全部测量指标所具有的信息，这样既减少了变量的个数又便于对实际问题的分析和研究。

#选择19个主成分

#pc.f19 <- principal(ybdf[,-79],nfactors = 19,rotate = "varimax" ,

scores = TRUE)

#选择20个主成分

pc.f20 <- principal(ybdf[,-79],nfactors = 20,rotate = "varimax" ,

scores = TRUE)

pc.f20$scores #主成分得分

pc.f20$weights #主成分得分的系数

表格

描述已自动生成

注：数据量过多，页面展示不完整。

在第一主成分的表达式中，月统筹金额\_MAX、月统筹金额\_AVG、月药品金额\_MAX、月药品金额\_AVG、医院\_统筹金\_MAX、医院\_统筹金\_AVG、医院\_药品\_MAX、医院\_药品\_AVG等指标的系数较大，可以看成是月统筹金额、XXXX的综合指标。

在第二主成分中，月就诊天数\_MAX、月就诊天数\_AVG、就诊次数\_SUM、月就诊次数\_MAX、月就诊次数\_AVG等指标的系数较大，可以看成是就诊次数方面的综合指标。

在第三主成分中，……。

#进行预测