**MBA 教育管理中心**

电话：029-82669064

传真：029-82668069

管理学院专业学位大作业

课程名称《商务大数据》

专业名称： MBA

班 级： 8146班 第5组

学号/姓名： 赵双喜3118426003

龙佳星3118426048董 璞3117580074

李 娜3118426053常 茵3118426016

谢舒文3118426014李 茜3118426026

赵丽轩3118426004

任课教师： 王 尧

提交时间： 2019年 7 月 6 日

# 目录

目录

[目录 2](#_Toc13144824)

[一、数据导入及处理 2](#_Toc13144825)

[二、查看各变量的影响 2](#_Toc13144826)

[1. 气缸数对油耗的影响 2](#_Toc13144827)

[2.发动机排量对油耗的影响 3](#_Toc13144828)

[3. 马力对油耗的影响 4](#_Toc13144829)

[4.车重对油耗的影响 5](#_Toc13144830)

[5.加速度对油耗的影响 6](#_Toc13144831)

[6.年份对油耗的影响 7](#_Toc13144832)

[7.产地对油耗的影响 8](#_Toc13144833)

[8. 合并绘图查看 9](#_Toc13144834)

[三、分类数据 10](#_Toc13144835)

[四、构建模型及预测 10](#_Toc13144836)

[五、反复执行统计差错率 15](#_Toc13144837)

[六、尝试通过去掉共线性优化 16](#_Toc13144838)

汽车油耗分析

# 一、数据导入及处理

library(readr)

dat0 <- read.csv('Auto.csv', header=T , encoding="UTF-8")#读文件

head(dat0)#看文件

median(dat0$mpg)#取中位数

dat0$mpg001[dat0$mpg>median(dat0$mpg)]=1#大于其中位数则令其为1

dat0$mpg001[dat0$mpg<=median(dat0$mpg)]=0

dat1=dat0[,-1]#去掉mpg列

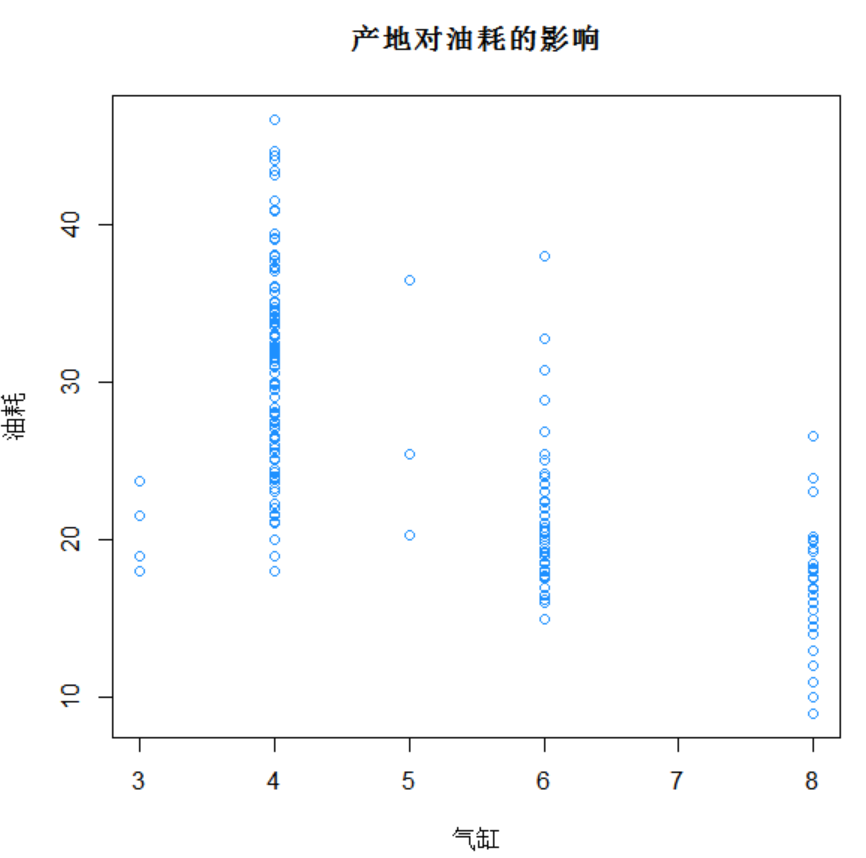
View(dat1)

# 二、查看各变量的影响

## 1. 气缸数对油耗的影响

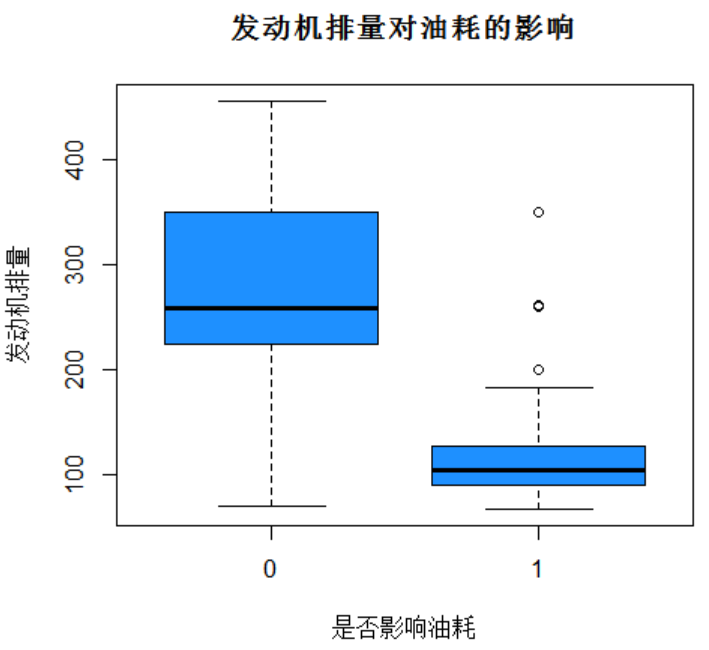
windows()

boxplot(cylinders~mpg001,data = dat1,col='dodgerblue',ylab = '气缸数',xlab = '是否影响油耗',ylim=c(4,8),main='气缸数对油耗的影响')



## 2.发动机排量对油耗的影响

boxplot(displacement~mpg001,data = dat1,col='dodgerblue',ylab = '发动机排量',xlab = '是否影响油耗',main='发动机排量对油耗的影响')



## 3. 马力对油耗的影响

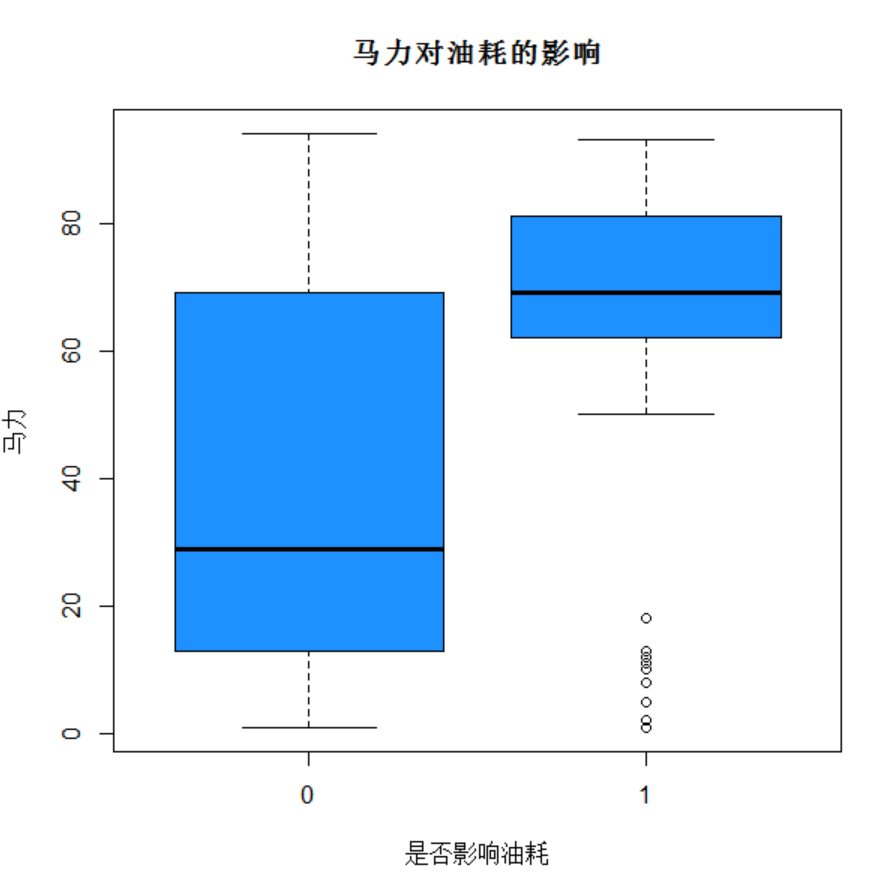
#马力对油耗的影响

class(dat1$horsepower)

#"numeric"

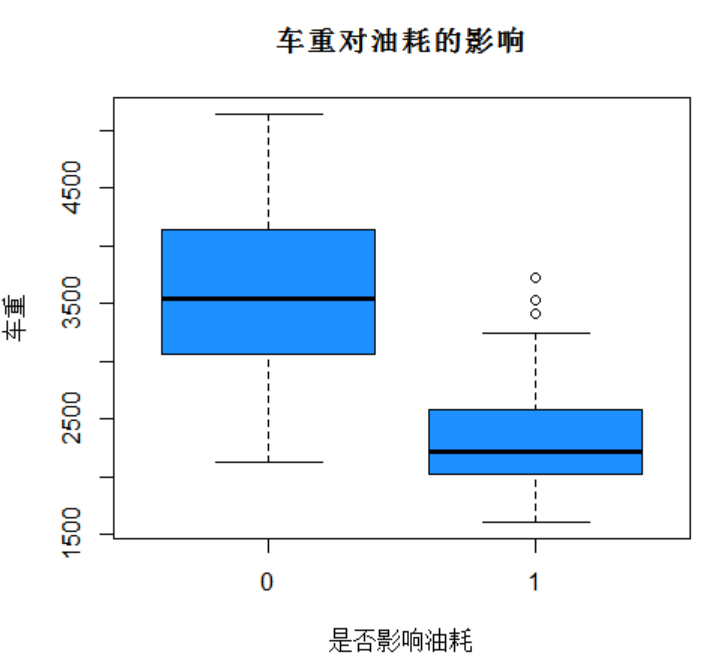
dat1$horsepower = as.numeric(dat1$horsepower) # 将horsepower的factor变量改为numeric变量

boxplot(horsepower~mpg001,data = dat1,col='dodgerblue',ylab = '马力',xlab = '是否影响油耗',main='马力对油耗的影响')



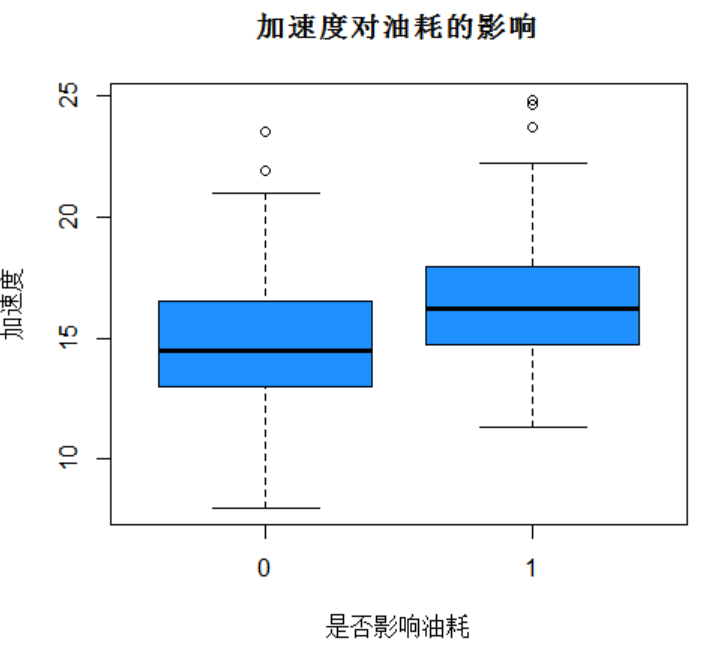
## 4.车重对油耗的影响

boxplot(weight~mpg001,data = dat1,col='dodgerblue',ylab = '车重',xlab = '是否影响油耗',main='车重对油耗的影响')



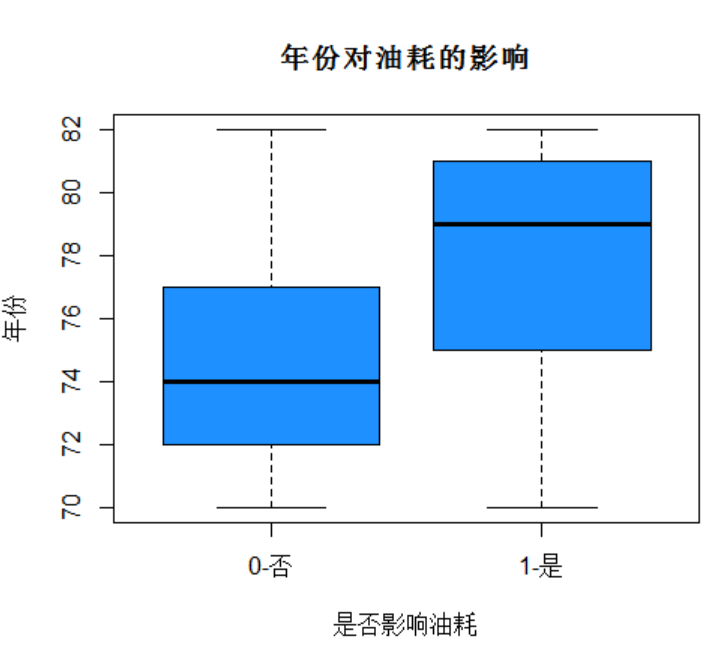
## 5.加速度对油耗的影响

boxplot(acceleration~mpg001,data = dat1,col='dodgerblue',ylab = '加速度',xlab = '是否影响油耗',main='加速度对油耗的影响')



## 6.年份对油耗的影响

boxplot(year~mpg001,data = dat1,col='dodgerblue',ylab = '年份',xlab = '是否影响油耗',main='年份对油耗的影响',xaxt="n")

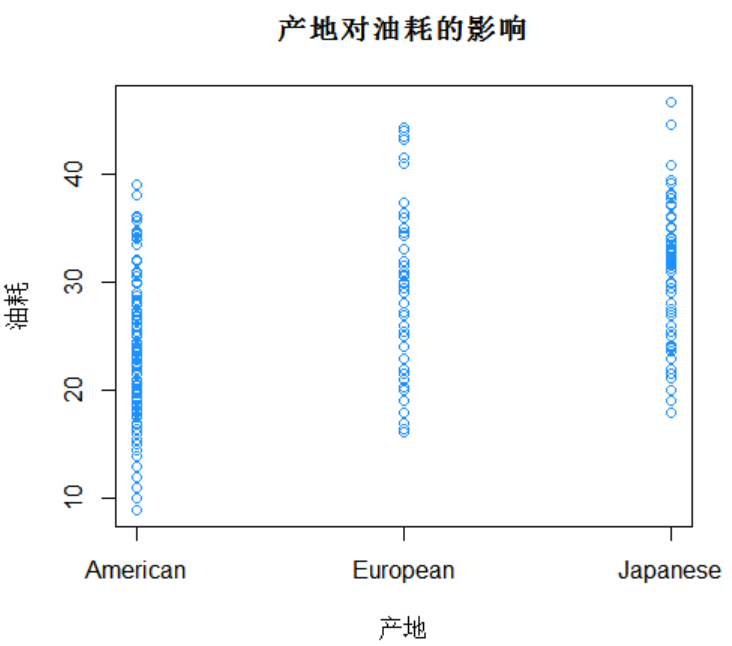
axis(1,at=c(1,2),labels=c("是","否")) 

## 7.产地对油耗的影响

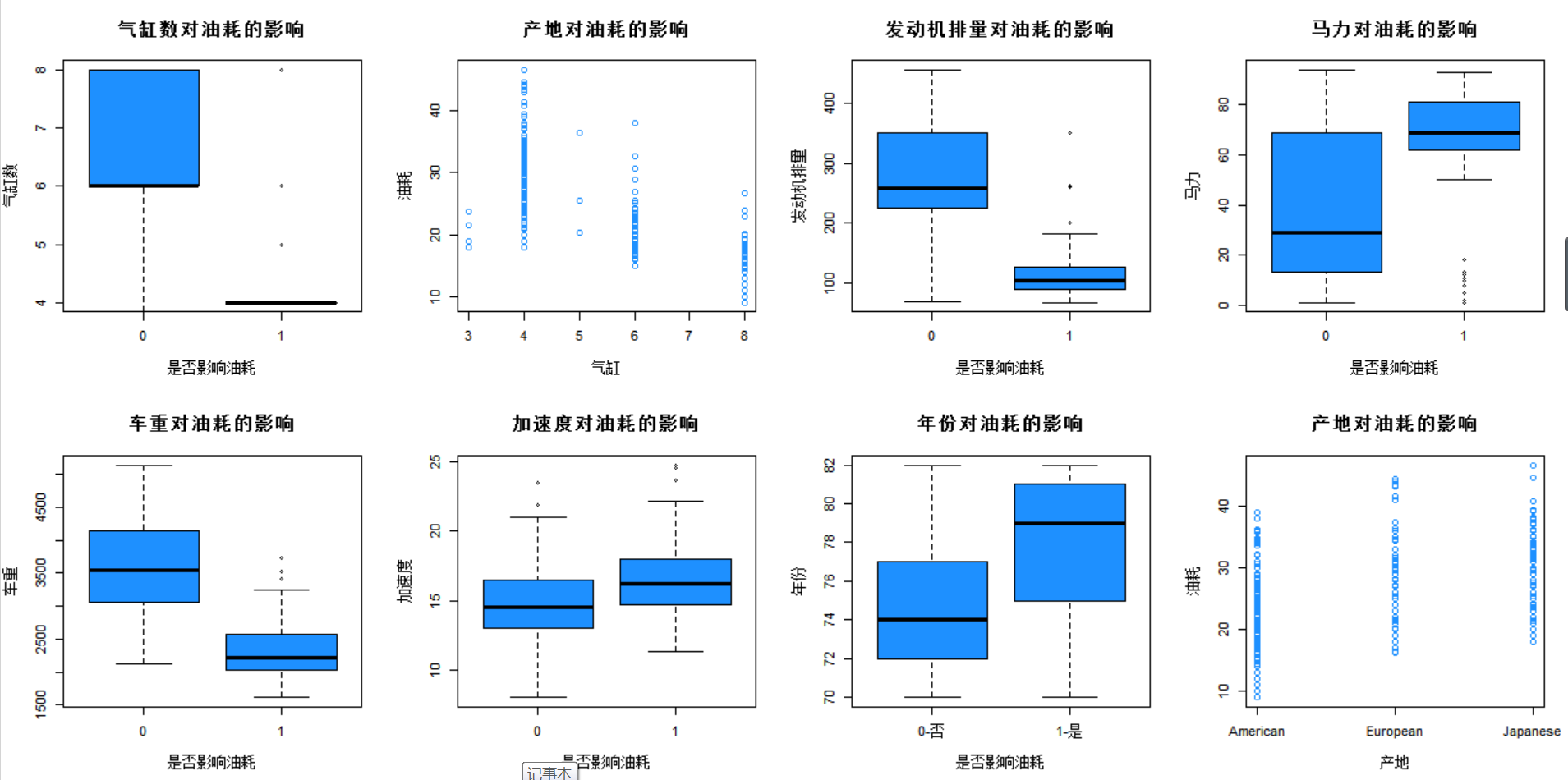
#产地对油耗的影响：

plot(dat0$origin,dat0$mpg,main='产地对油耗的影响',xaxt="n",xlab='产地',ylab='油耗')

axis(1,at=c(1,2,3),labels=c("American","European","Japanese"))



## 8. 合并绘图查看



# 三、分类数据

## 60% of the sample size

smp\_size <- floor(0.60 \* nrow(dat1))

## set the seed to make your partition reproducible

set.seed(123)

train\_ind <- sample(seq\_len(nrow(dat1)), size = smp\_size)

train <-dat1[train\_ind, ]

test <- dat1[-train\_ind, ]

# 四、构建模型及预测

train=train[,-8]#去掉name列

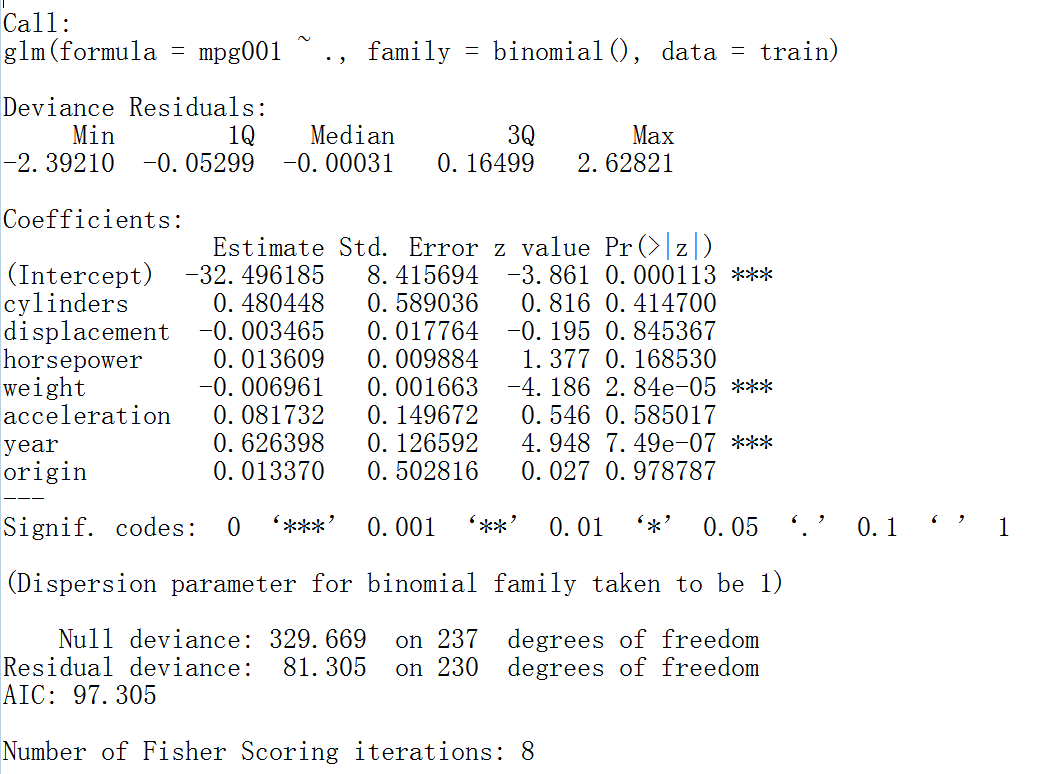
test=test[,-8]#去掉name列

lm1=glm(mpg001~.,data=train,family=binomial())

sink("summary.txt",append=FALSE,split=TRUE)#控制台不好复制，存下来

print(summary(lm1))

sink()



Yhat估计出来他是一个概率，那么我们要对概率有一个指标的估计，这到底说这个概率大约一个多少值，它取值为1，然后其他的就取为0。

如果算出来它的概率大于0.5，我们就确认为1，否则为0。这样的方式是比较标准的方式，那么第二个方式是取的是所有数据指标的一个平均值，作为这个概率的一个阈值，我一般建议大家就做第一种就好了，0.5是一个比较标准的做法。

第二种，就是说他取了平均值，对这种反正有点启发式，都是这种概率的做法，都比较启发是吧，我觉得建议大家可以在0.50，0.6，0.4，反正这几个里面。都可以算一算，一般正常来说，大部分的文章里面的建议0.5。

Yhat=predict(lm1,newdata=test,type='response')#type是指选择预测的类型，由于是二分类变量，所以选择reponse，表示输出结果预测响应变量为1的概率

ypre1=1\*(Yhat>0.5)

ypre2=1\*(Yhat>mean(test$mpg001))

#install.packages('pROC')

library(pROC)

windows()

plot.roc(test$mpg001,Yhat,col='red',lwd=2,xaxs='i',yaxs='i')

#ROC曲线

#family=gaussian(线性回归)

#family=bionmial(逻辑回归)

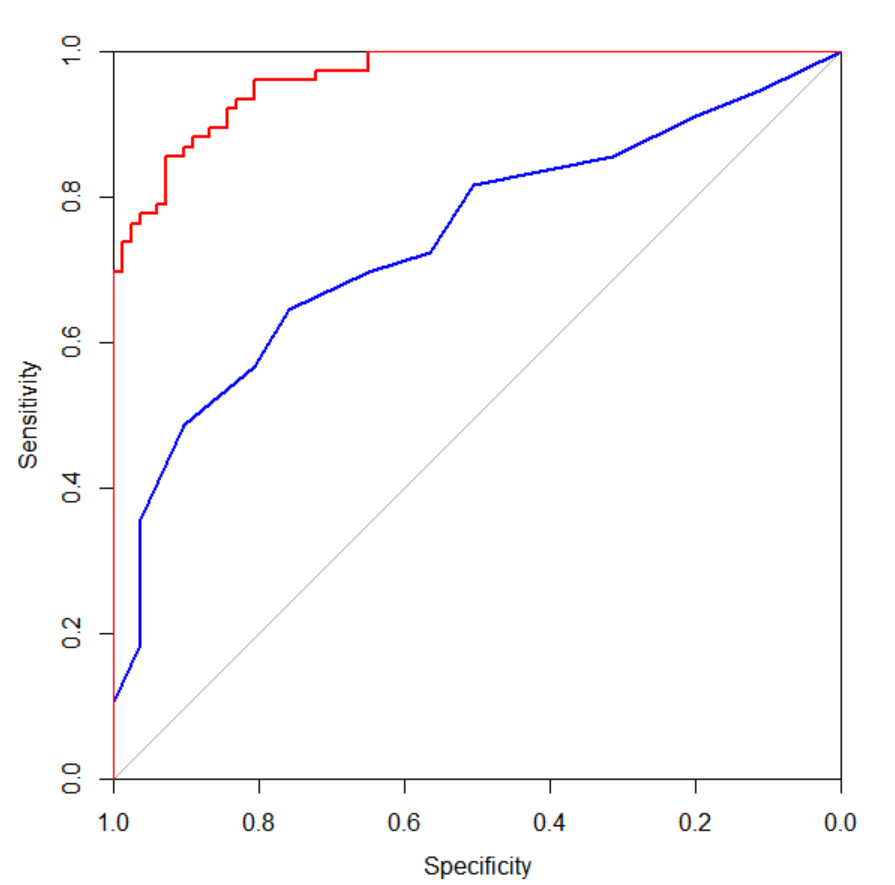
lm2=glm(mpg001~year,data=train,family=binomial())

Yhat2=predict(lm2,newdata=test,type='response')

windows()

plot.roc(test$mpg001,Yhat2,col='blue',lwd=2,xaxs='i',yaxs='i')

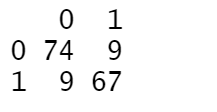
lines.roc(test$mpg001,Yhat,col='red',lwd=2)

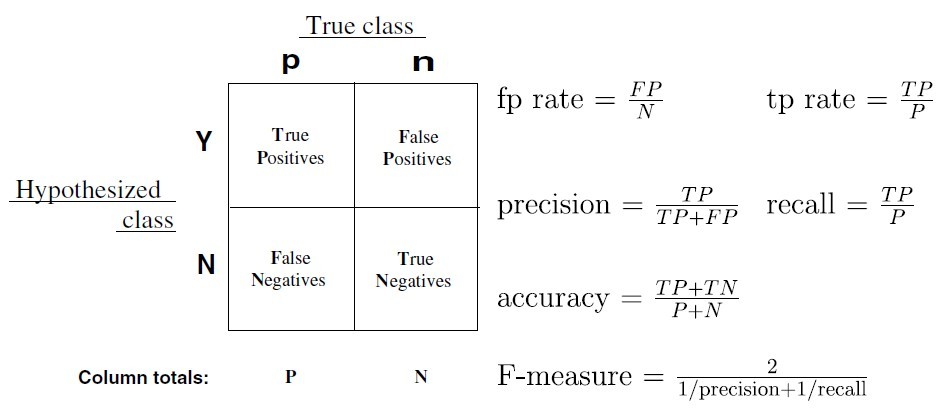


ypre1=1\*(Yhat>0.5)

test$mpgpreict1 <- ypre1

table(test$mpg001,test$mpgpreict1)#查看预测情况





召回率(Recall) = TP/P = TPR

准确率(Precision) = TP/（TP+FP）

一共正确的个数是74+67，一共错误的个数是9+9，那么误差率也就是错分率是（74+67）/(74+67+9+9)

分类问题的错分率：

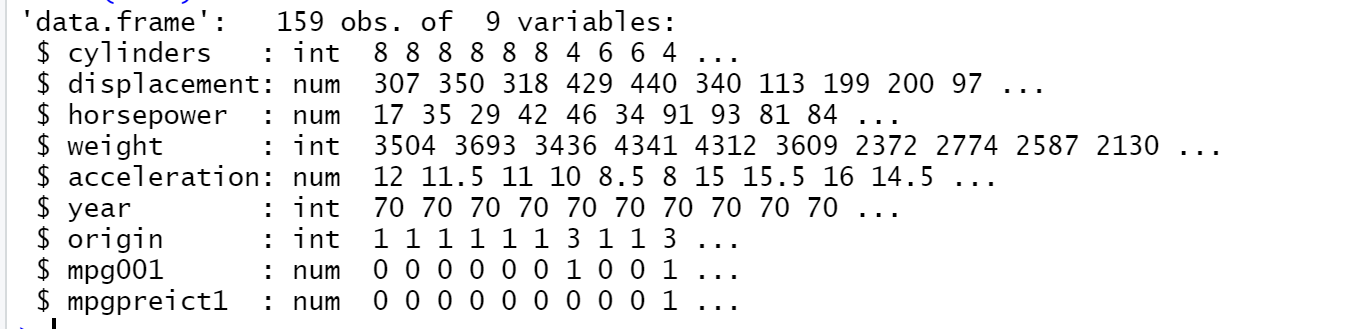
在模型里面，我们要分成训练集测试集，那么测试集啊，我们有一个X，对吧，也有他的Y。假定这个Y，我们是不知道的，就是说，比如说就是上帝知道，对吧，那么这个时候我们决定X之后，用我们的这个模型，算出来，用predict函数把一个说清楚，X输进去，那么就得到一个Y。那么这个Y，然后再给你真实的这个Y作一个对比，如果他是1，我算出来也是1的话，那么我们算对了，如果他是0，我算出来是0的话，那么我也算对了。但如果他是1我算成0，或者它是0，我算了1，那就是我算错了，我就统计一下这个测试题上有多少个个数，算错了。用算错的个数除以总的样本数，就是误差率。

计算差错率：

error=1-sum(test$mpgpreict1==test$mpg001)/length(test$mpg001)

#sum(test$mpgpreict1==test$mpg001)/length(test$mpg001) 这个是正确的比率，然后1减去它就是错误的比率

str(test)#查看有多少行。

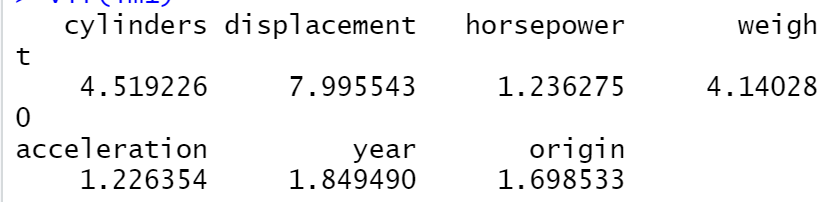


# 五、反复执行统计差错率

|  |  |
| --- | --- |
| 序号 | 差错率 |
| 1 | 0.113208 |
| 2 | 0.075472 |
| 3 | 0.157233 |
| 4 | 0.081761 |
| 5 | 0.106918 |
| 6 | 0.106918 |
| 7 | 0.119497 |
| 8 | 0.119497 |
| 9 | 0.062893 |
| 10 | 0.119497 |
| 11 | 0.106918 |
| 12 | 0.081761 |
| 13 | 0.08805 |
| 14 | 0.081761 |
| 15 | 0.106918 |
| 16 | 0.09434 |
| 17 | 0.081761 |
| 18 | 0.056604 |
| 19 | 0.08805 |
| 20 | 0.069182 |
| 平均值 | 0.095912 |

# 六、尝试通过去掉共线性优化

vif(lm1)



去掉排量影响

#以下代码复制反复运行，查看最后一个结果：

smp\_size <- floor(0.60 \* nrow(dat1))#分比例

train\_ind <- sample(seq\_len(nrow(dat1)), size = smp\_size)#取60%

train <-dat1[train\_ind, ]#存为训练集

test <- dat1[-train\_ind, ]#除去训练集即为测试集

train=train[,-8]#去掉name列

test=test[,-8]#去掉name列

train=train[,-2]#去掉displacement列

test=test[,-2]#去掉displacement列

lm3=glm(mpg001~.,data=train,family=binomial())#构建模型

Yhat=predict(lm3,newdata=test,type='response')#type是指选择预测的类型，由于是二分类变量，所以选择reponse，表示输出结果预测响应变量为1的概率

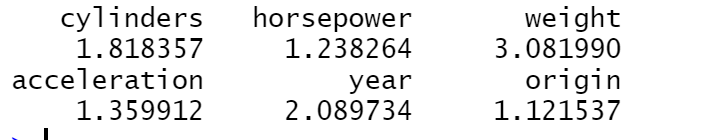
ypre3=1\*(Yhat>0.5)

test$mpgpreict1 <- ypre3#表里增加一列。

error=1-sum(test$mpgpreict1==test$mpg001)/length(test$mpg001)#计算错分率

print(error)

vif(lm3)



|  |  |  |
| --- | --- | --- |
| 序号 | 差错率全部变量 | 差错率去掉排量 |
| 1 | 0.1132075 | 0.08805031 |
| 2 | 0.0754717 | 0.1194969 |
| 3 | 0.1572327 | 0.0754717 |
| 4 | 0.08176101 | 0.1383648 |
| 5 | 0.1069182 | 0.09433962 |
| 6 | 0.1069182 | 0.09433962 |
| 7 | 0.1194969 | 0.09433962 |
| 8 | 0.1194969 | 0.1006289 |
| 9 | 0.06289308 | 0.06289308 |
| 10 | 0.1194969 | 0.1006289 |
| 11 | 0.1069182 | 0.1006289 |
| 12 | 0.08176101 | 0.09433962 |
| 13 | 0.08805031 | 0.1257862 |
| 14 | 0.08176101 | 0.1257862 |
| 15 | 0.1069182 | 0.08176101 |
| 16 | 0.09433962 | 0.09433962 |
| 17 | 0.08176101 | 0.0754717 |
| 18 | 0.05660377 | 0.05031447 |
| 19 | 0.08805031 | 0.09433962 |
| 20 | 0.06918239 | 0.1006289 |
| 平均值 | 0.095911946 | 0.095597485 |
| 差值 |  | 0.000314461 |

两者偏差非常小，共线性影响不大。

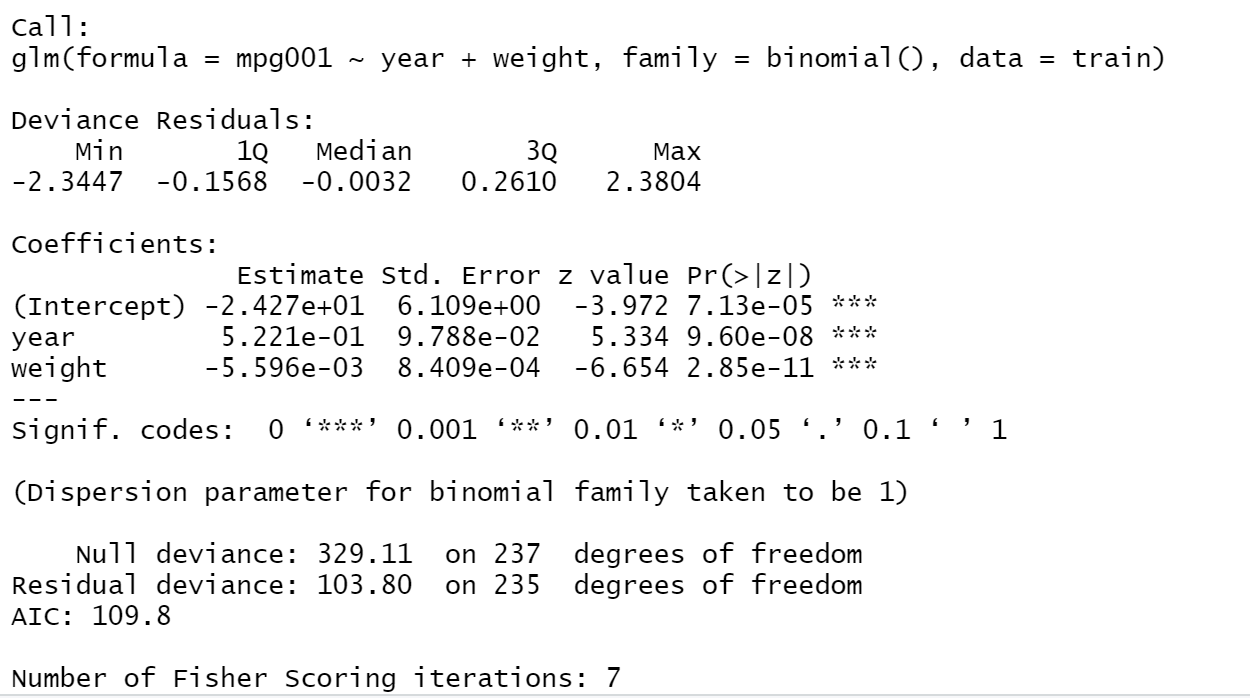
测试采用重量、年

只做year+weight，及P<0.05的

lm4=glm(mpg001~year+weight,data=train,family=binomial())#只取一个年份做模型看下

Yhat4=predict(lm4,newdata=test,type='response')

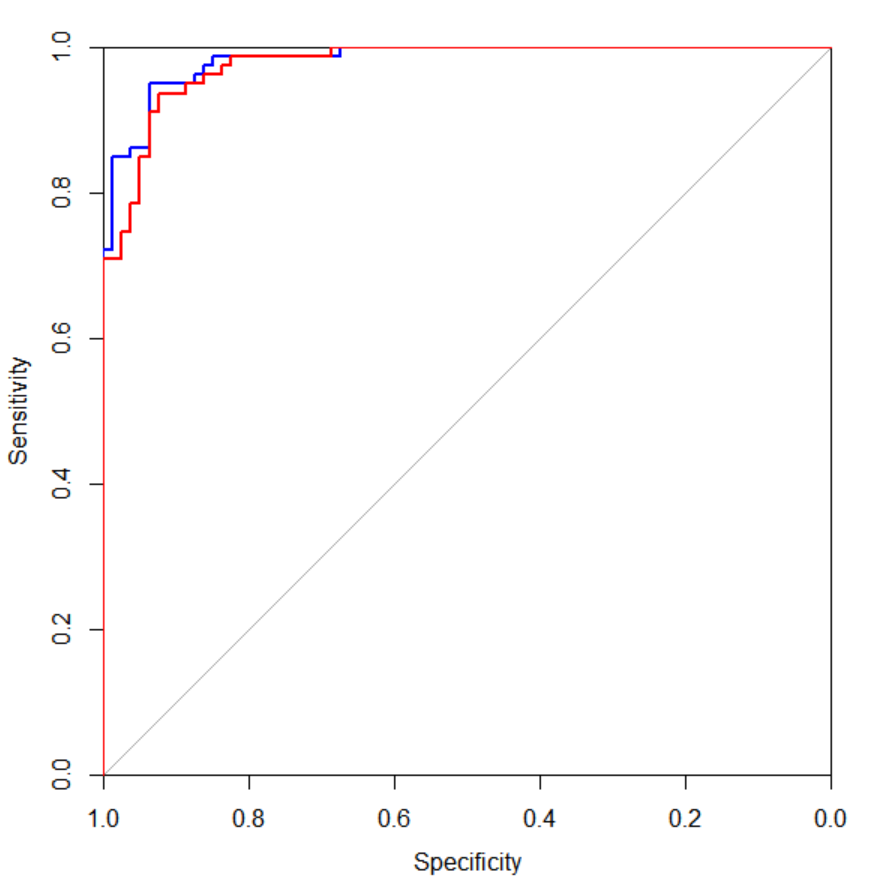
(summary(lm4))



windows()

plot.roc(test$mpg001,Yhat4,col='blue',lwd=2,xaxs='i',yaxs='i')

lines.roc(test$mpg001,Yhat,col='red',lwd=2)



数据量太小，训练结果在当前水平非常难以提高了。

