STA403-1 第一次作业

金科 201756010

### 第一周任务：P38 第8题

1. 用read.csv()函数将该数据读进R，将载入的数据命名为college

college <- read.csv("ISLR\_data/College.csv", header = T)

1. 用fix()函数观察数据

rownames(college) = college[, 1]  
fix(college)

row.names不是数据列，而是每个记录的名字

college = college[, -1]  
fix(college)

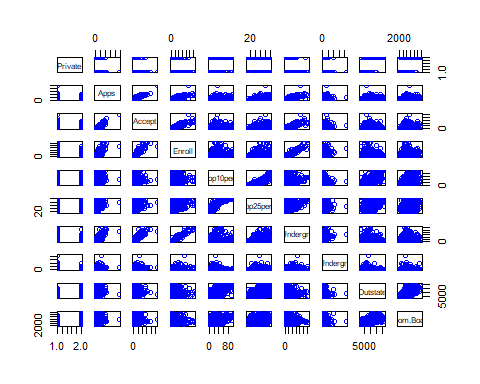
1. 汇总统计信息

summary(college)

## Private Apps Accept Enroll Top10perc   
## No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00   
## Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00   
## Median : 1558 Median : 1110 Median : 434 Median :23.00   
## Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00   
## Max. :48094 Max. :26330 Max. :6392 Max. :96.00   
## Top25perc F.Undergrad P.Undergrad Outstate   
## Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340   
## 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320   
## Median : 54.0 Median : 1707 Median : 353.0 Median : 9990   
## Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441   
## 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925   
## Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700   
## Room.Board Books Personal PhD   
## Min. :1780 Min. : 96.0 Min. : 250 Min. : 8.00   
## 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00   
## Median :4200 Median : 500.0 Median :1200 Median : 75.00   
## Mean :4358 Mean : 549.4 Mean :1341 Mean : 72.66   
## 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00   
## Max. :8124 Max. :2340.0 Max. :6800 Max. :103.00   
## Terminal S.F.Ratio perc.alumni Expend   
## Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186   
## 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751   
## Median : 82.0 Median :13.60 Median :21.00 Median : 8377   
## Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660   
## 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830   
## Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233   
## Grad.Rate   
## Min. : 10.00   
## 1st Qu.: 53.00   
## Median : 65.00   
## Mean : 65.46   
## 3rd Qu.: 78.00   
## Max. :118.00

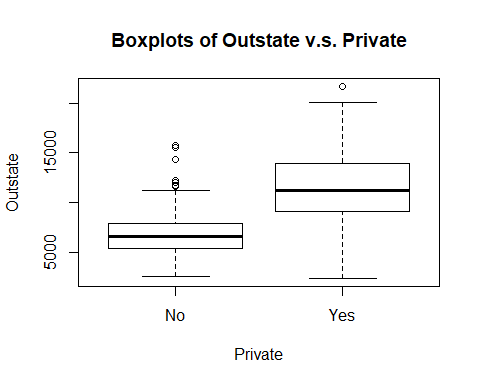
对前十列变量生成散点图矩阵

pairs(college[, 1:10], col = "blue")



生成Outstate和Private变量的沿边箱线图

plot(college$Private, college$Outstate,   
 xlab = "Private", ylab = "Outstate",   
 main = "Boxplots of Outstate v.s. Private")

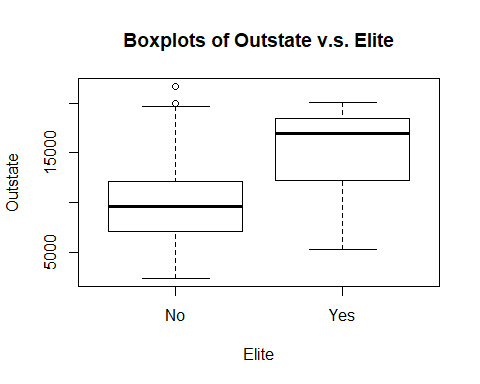


排名在前10%顶尖高中的情况分组创建新变量

Elite = rep("No", nrow(college)) # 初始化Elite变量  
Elite[college$Top10perc > 50] = "Yes" # 根据要求分组  
Elite = as.factor(Elite) # 转换为因子型  
college = data.frame(college, Elite) # 合并生成新的数据框  
summary(Elite)

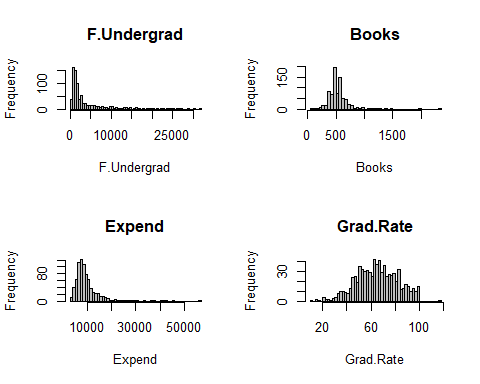
## No Yes   
## 699 78

plot(college$Elite, college$Outstate,   
 xlab = "Elite", ylab = "Outstate",   
 main = "Boxplots of Outstate v.s. Elite")



直方图

attach(college)  
par(mfrow = c(2,2))  
hist(F.Undergrad, breaks = 50, col = "gray"，main = "F.Undergrad")  
hist(Books, breaks = 50, col = "gray", main = "Books")  
hist(Expend, breaks = 50, col = "gray", main = "Expend")  
hist(Grad.Rate, breaks = 50, col = "gray", main = "Grad.Rate")



继续探索数据可以发现，**MIT**校园里有着最多的来自前10%顶尖高中的学生，**卡泽诺维亚学院**的毕业率最高。

row.names(college)[which.max(Top10perc)]

## [1] "Massachusetts Institute of Technology"

row.names(college)[which.max(Grad.Rate)]

## [1] "Cazenovia College"

### 第二周任务：P87 第15题

1. 简单回归

library(MASS)

lmp <- function(x) {  
 lm <- lm(Boston$crim ~ x)  
 f <- summary(lm)$fstatistic  
 p <- pf(f[1],f[2],f[3], lower.tail=F)  
 names(p) <- NULL  
 return(p)  
}  
  
results <- apply(Boston[, -1], 2, lmp)  
names(results) <- paste(names(Boston)[1],   
 names(Boston)[2:14],   
 sep = " ~ ")  
results[results>=0.00001]

## crim ~ chas   
## 0.2094345

由结果可以看出，除了chas以外，其他预测变量与人均犯罪率crim的模型都非常显著（）。

1. 多元回归

lm <- lm(crim ~., data = Boston)  
summary(lm)

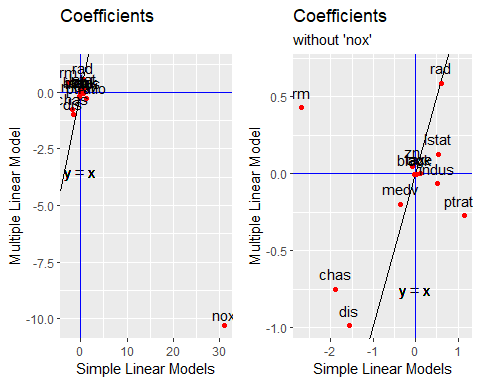
##   
## Call:  
## lm(formula = crim ~ ., data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.924 -2.120 -0.353 1.019 75.051   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.033228 7.234903 2.354 0.018949 \*   
## zn 0.044855 0.018734 2.394 0.017025 \*   
## indus -0.063855 0.083407 -0.766 0.444294   
## chas -0.749134 1.180147 -0.635 0.525867   
## nox -10.313535 5.275536 -1.955 0.051152 .   
## rm 0.430131 0.612830 0.702 0.483089   
## age 0.001452 0.017925 0.081 0.935488   
## dis -0.987176 0.281817 -3.503 0.000502 \*\*\*  
## rad 0.588209 0.088049 6.680 6.46e-11 \*\*\*  
## tax -0.003780 0.005156 -0.733 0.463793   
## ptratio -0.271081 0.186450 -1.454 0.146611   
## black -0.007538 0.003673 -2.052 0.040702 \*   
## lstat 0.126211 0.075725 1.667 0.096208 .   
## medv -0.198887 0.060516 -3.287 0.001087 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.439 on 492 degrees of freedom  
## Multiple R-squared: 0.454, Adjusted R-squared: 0.4396   
## F-statistic: 31.47 on 13 and 492 DF, p-value: < 2.2e-16

如上结果所示，在0.05的显著性水平下，变量zn、dis、rad、black、medv可以拒绝零假设。

# 简单回归系数  
coef\_1 <- apply(Boston[, -1], 2,   
 function(x){coefficients(lm(Boston$crim~x))[2]})  
# 多元回归系数  
coef\_2 <- coefficients(lm)  
coef\_2 <- as.data.frame(coef\_2)[-1,]  
  
# 合并数据框  
coef <- data.frame(coef\_1, coef\_2)

绘制散点图，比较回归系数

library(ggplot2)  
  
p1 <- ggplot(coef, aes(x = coef\_1, y = coef\_2)) +   
 geom\_abline(intercept = 0, slope = 1) +   
 geom\_text(aes(0, -3.5, label = "y = x")) +   
 geom\_vline(xintercept = 0, color = "blue") +   
 geom\_hline(yintercept = 0, color = "blue") +  
 geom\_point(color = "red") +   
 geom\_text(aes(y = coef\_2 + .5, label = names(Boston)[-1])) +   
 labs(x = "Simple Linear Models",   
 y = "Multiple Linear Model",  
 title = "Coefficients",   
 subtitle = "")  
  
p2 <- ggplot(coef[-4,], aes(x = coef\_1, y = coef\_2)) +   
 geom\_abline(intercept = 0, slope = 1) +   
 geom\_text(aes(0, -.75, label = "y = x")) +   
 geom\_vline(xintercept = 0, color = "blue") +   
 geom\_hline(yintercept = 0, color = "blue") +  
 geom\_point(color = "red") +   
 geom\_text(aes(y = coef\_2 + .1, label = names(Boston)[c(-1,-5)])) +   
 labs(x = "Simple Linear Models",   
 y = "Multiple Linear Model",   
 title = "Coefficients",   
 subtitle = "without 'nox'")  
  
# 合并显示  
cowplot::plot\_grid(p1, p2)



左图可以看出，变量nox在两种模型中的回归系数相距甚远；如右图，变量rm、indus和ptratio位于二、四象限，说明在两个模型中它们的回归系数正负符号不同，拟合情况相差较大。

1. 对每个预测变量**分别**进行多项式回归

coefp <- function (x) {  
 lm <- lm(Boston$crim ~ x+I(x^2)+I(x^3))  
 cf <- summary(lm)$coefficients  
 return(cf[-1,4])  
}  
res\_ploy <- apply(Boston[, -1], 2, coefp)  
res\_ploy\_dataframe <- as.data.frame(res\_ploy)  
t(res\_ploy\_dataframe)

## x I(x^2) I(x^3)  
## zn 2.612296e-03 9.375050e-02 2.295386e-01  
## indus 5.297064e-05 3.420187e-10 1.196405e-12  
## chas 2.094345e-01 2.094345e-01 2.094345e-01  
## nox 2.758372e-13 6.811300e-15 6.961110e-16  
## rm 2.117564e-01 3.641094e-01 5.085751e-01  
## age 1.426608e-01 4.737733e-02 6.679915e-03  
## dis 6.374792e-18 4.941214e-12 1.088832e-08  
## rad 6.234175e-01 6.130099e-01 4.823138e-01  
## tax 1.097075e-01 1.374682e-01 2.438507e-01  
## ptratio 3.028663e-03 4.119552e-03 6.300514e-03  
## black 1.385871e-01 4.741751e-01 5.436172e-01  
## lstat 3.345300e-01 6.458736e-02 1.298906e-01  
## medv 2.637707e-28 3.260523e-18 1.046510e-12

回归结果表明，响应变量与预测变量存在非线性关系，尤其在变量zn、indus、nox、dis、ptratio、medv上比较明显。