**代码清单8-1**

> fit<-lm(weights = height,data=women)

> summary(fit)

Call:

lm(formula = weight ~ height, data = women)

Residuals:

Min 1Q Median 3Q Max

-1.7333 -1.1333 -0.3833 0.7417 3.1167

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -87.51667 5.93694 -14.74 1.71e-09 \*\*\*

height 3.45000 0.09114 37.85 1.09e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.525 on 13 degrees of freedom

Multiple R-squared: 0.991, Adjusted R-squared: 0.9903

F-statistic: 1433 on 1 and 13 DF, p-value: 1.091e-14

解释：

* Residuals：残差项的最小值是-1.7333，最大值是3.1167，1分位点是-1.1333，中值是-0.3833，3分位点是0.7417。
* Coefficients:截距的估计值为-87.51667，标准误差为5.93694，t值为-14.74，p值为1.71e-09。height的系数估计值为3.45000，标准误差为0.09114，t值为37.85，p值为1.09e-14，它们的P值均远远小于0.05，通过显著性检验。
* 拟合优度Multiple R-squared为0.991接近与1，它的拟合程度很好。
* F-statistic为F检验，用于判断方程整体的显著性检验，其P值为1.091e-14远远小于0.05，其显著性很好。
* > fitted(fit)

1 2 3 4 5 6 7 8

112.5833 116.0333 119.4833 122.9333 126.3833 129.8333 133.2833 136.7333

9 10 11 12 13 14 15

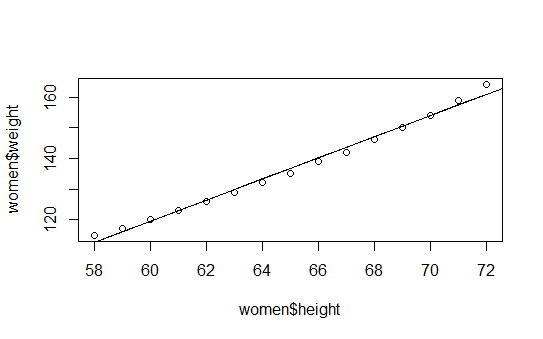
140.1833 143.6333 147.0833 150.5333 153.9833 157.4333 160.8833

解释：

获取回归模型上对应的点

plot(women$height,women$weight)

abline(fit)



圆圈为原始值，直线为拟合的直线

**代码清单8-2**

> fit2<-lm(weight ~height+I(height^2),data=women)

> summary(fit2)

Call:

lm(formula = weight ~ height + I(height^2), data = women)

Residuals:

Min 1Q Median 3Q Max

-0.50941 -0.29611 -0.00941 0.28615 0.59706

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 261.87818 25.19677 10.393 2.36e-07 \*\*\*

height -7.34832 0.77769 -9.449 6.58e-07 \*\*\*

I(height^2) 0.08306 0.00598 13.891 9.32e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3841 on 12 degrees of freedom

Multiple R-squared: 0.9995, Adjusted R-squared: 0.9994

F-statistic: 1.139e+04 on 2 and 12 DF, p-value: < 2.2e-16

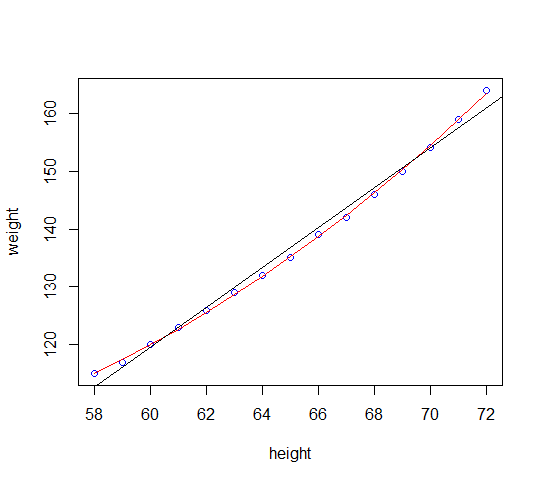
解释：

* 拟合优度Multiple R-squared为0.9995，方差解释率增加到了99.9%，比简单线性回归的0.991好

> plot(women$height,women$weight,xlab = "height",ylab ="weight" )

> lines(women$height,fitted(fit2))

> abline(fit)



圆圈为原始值，直线为线性拟合，曲线为多项式拟合

**代码清单8-3**

> states <- as.data.frame(state.x77[, c("Murder", "Population","Illiteracy","Income","Frost")])

> cor(states)

Murder Population Illiteracy Income Frost

Murder 1.0000000 0.3436428 0.7029752 -0.2300776 -0.5388834

Population 0.3436428 1.0000000 0.1076224 0.2082276 -0.3321525

Illiteracy 0.7029752 0.1076224 1.0000000 -0.4370752 -0.6719470

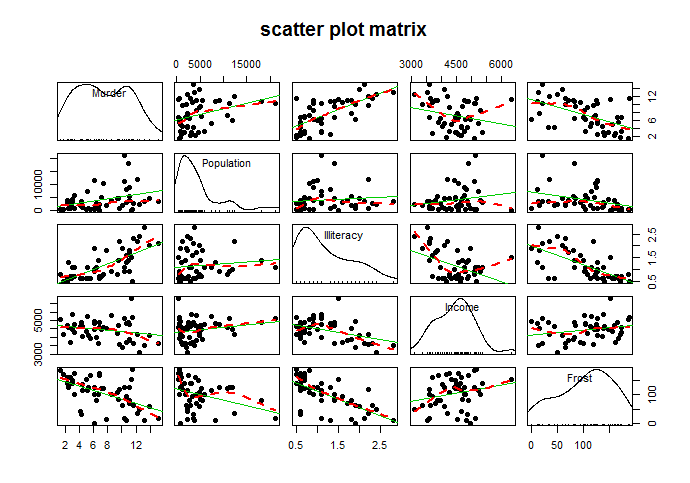
Income -0.2300776 0.2082276 -0.4370752 1.0000000 0.2262822

Frost -0.5388834 -0.3321525 -0.6719470 0.2262822 1.0000000

解释：协方差矩阵，判断两两变量之间的相关性，越接近1越正相关，越接近-1越负相关，谋杀率与文盲率有很强的相关性

> library(car)

> scatterplotMatrix(states,spread=FALSE,smoother.args=list(lty=2),pch=19,main="scatter plot matrix")



在非对角线区域绘制了变量的散点图和loess拟合曲线以及线性拟合曲线，谋杀率随着人口和文盲率上升而上升，随着收入和结霜天数增加而下降，收入水平越高谋杀率越低

**代码清单8-4**

> states <- as.data.frame(state.x77[, c("Murder", "Population","Illiteracy","Income","Frost")])

> fit3<-lm(Murder~Population+Illiteracy+Income+Frost,data=states)

> summary(fit3)

Call:

lm(formula = Murder ~ Population + Illiteracy + Income + Frost,

data = states)

Residuals:

Min 1Q Median 3Q Max

-4.7960 -1.6495 -0.0811 1.4815 7.6210

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.235e+00 3.866e+00 0.319 0.7510

Population 2.237e-04 9.052e-05 2.471 0.0173 \*

Illiteracy 4.143e+00 8.744e-01 4.738 2.19e-05 \*\*\*

Income 6.442e-05 6.837e-04 0.094 0.9253

Frost 5.813e-04 1.005e-02 0.058 0.9541

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.535 on 45 degrees of freedom

Multiple R-squared: 0.567, Adjusted R-squared: 0.5285

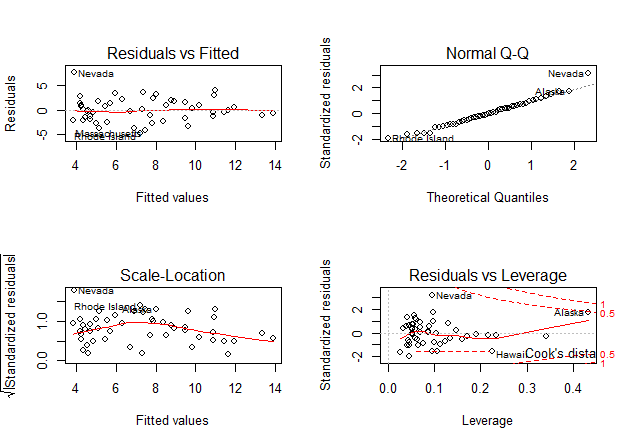
F-statistic: 14.73 on 4 and 45 DF, p-value: 9.133e-08

解释：

* Residuals：残差项的最小值是-4.7960，最大值是7.6210，1分位点是-1.6495，中值是-0.0811，3分位点是1.4815。
* Coefficients:截距的估计值为1.235，标准误差为3.866e，t值为0.319，p值为0.7510；Population 的系数估计值为2.237e-04，标准误差为9.052e-05，t值为2.471，p值为0.0173；Illiteracy 的系数估计值为4.143，标准误差为8.744e-01，t值为4.738，p值为2.19e-05；Income 的系数估计值为6.442e-05，标准误差为6.837e-04，t值为 0.094，p值为0.9253；Frost 的系数估计值为5.813e-04，标准误差为1.005e-02，t值为0.058，p值为0.9541。只有Illiteracy 的p值小于0.05，显著性水平高，其次是Population ，其他的显著性水平都比较低。
* 拟合优度Multiple R-squared为0.567，只能解释56.7%的方差。
* F-statistic为F检验，用于判断方程整体的显著性检验,它的值大于1，且P值为9.133e-08小于0.05，其显著性很好。

> par(mfrow=c(2,2))

> plot(fit3)



* 左上是残差与拟合值分布图，由残差图看不出明显规律，可以认为残差之间基本没有相关性
* 右上图表示如果残差呈现正态分布，那么图上的点落在45度角的直线上，由图可知只有部分点没有落在45度线上，也就是说残差基本呈现正太分布
* 左下与左上类似，由图可知，残差基本服从同方差性
* 右下是残差杠杆图，nevada有较高的学生化残差，超过2，可以认为是离群点，而alaska即有高残差杠杆也比较高，是要需要注意的点。

residplot <- function(fit, nbreaks=10) {

z <- rstudent(fit)

hist(z, breaks=nbreaks, freq=FALSE,

xlab="Studentized Residual",

main="Distribution of Errors")

rug(jitter(z), col="brown")

curve(dnorm(x, mean=mean(z), sd=sd(z)),

add=TRUE, col="blue", lwd=2)

lines(density(z)$x, density(z)$y,

col="red", lwd=2, lty=2)

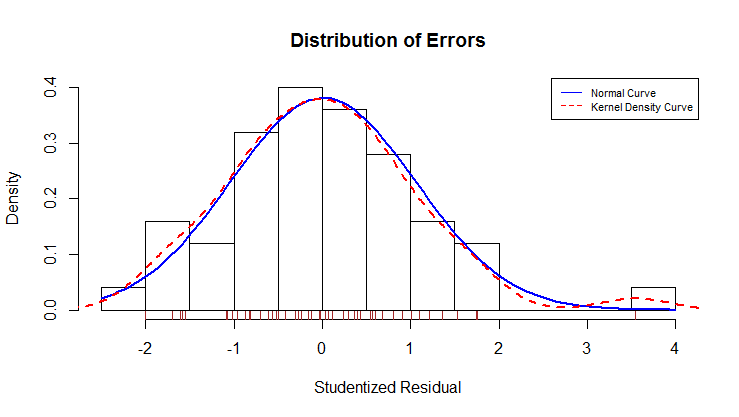
legend("topright",

legend = c( "Normal Curve", "Kernel Density Curve"),

lty=1:2, col=c("blue","red"), cex=.7)

}

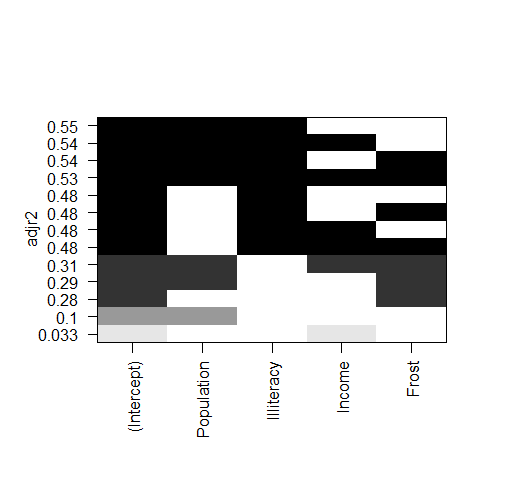
residplot(fit3)



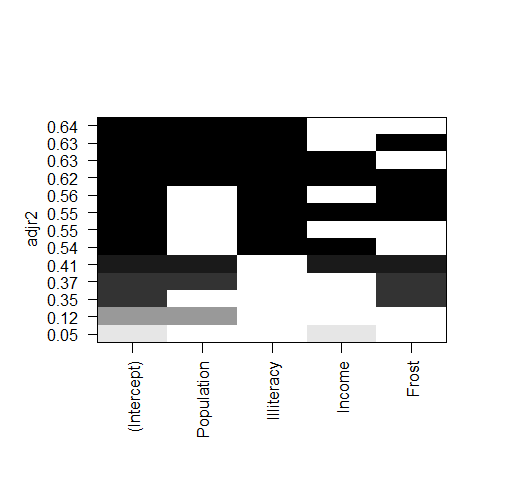
学生化残差柱状图，并添加了正太曲线，核密度曲线以及轴须图，误差很好的服从了正态分布，有一个明显的离群点。

全子集回归：

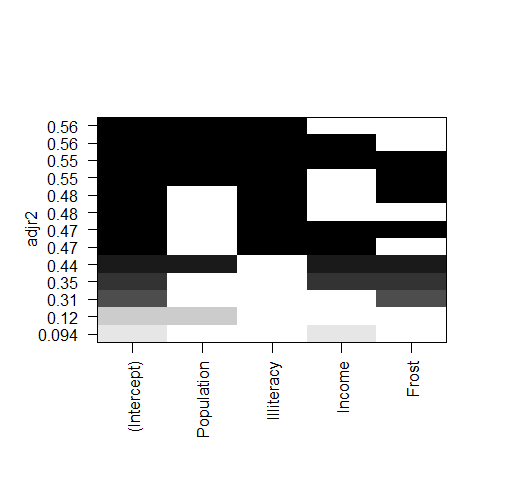
原始数据



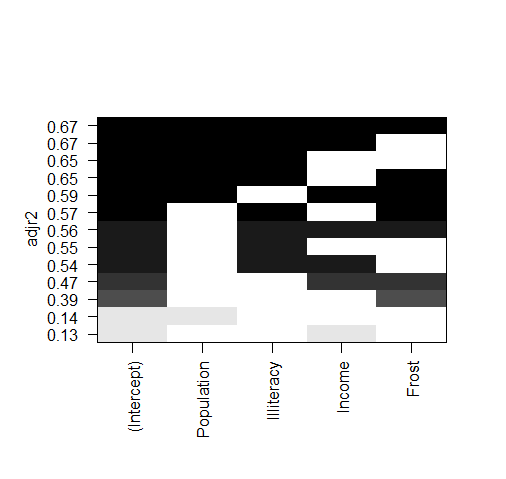
删了离群点Nevada



只删除alaska（有高残差，高杠杆）



删除alaska和Nevada



交叉验证

shrinkage <- function(fit, k = 10) {

if(!require(bootstrap)){

install.packages("bootstrap")

}

require(bootstrap)

# define functions

theta.fit <- function(x, y) {

lsfit(x, y)

}

theta.predict <- function(fit, x) {

cbind(1, x) %\*% fit$coef

}

# matrix of predictors

x <- fit$model[, 2:ncol(fit$model)]

# vector of predicted values

y <- fit$model[, 1]

results <- crossval(x, y, theta.fit, theta.predict, ngroup = k)

r2 <- cor(y, fit$fitted.values)^2

r2cv <- cor(y, results$cv.fit)^2

cat("Original R-square =", r2, "\n")

cat(k, "Fold Cross-Validated R-square =", r2cv, "\n")

cat("Change =", r2 - r2cv, "\n")

}

fit7<-lm(Murder~Population+Illiteracy+Income+Frost,data=states)

shrinkage(fit7)

fit8<-lm(Murder~Population+Illiteracy,data=states)

shrinkage(fit8)

fit9<-lm(Murder~Population+Illiteracy,data=statesDel)

shrinkage(fit9)

> fit7<-lm(Murder~Population+Illiteracy+Income+Frost,data=states)>

shrinkage(fit7)

Original R-square = 0.5669502

10 Fold Cross-Validated R-square = 0.4562084

Change = 0.1107418

> fit8<-lm(Murder~Population+Illiteracy,data=states)>

shrinkage(fit8)

Original R-square = 0.5668327

10 Fold Cross-Validated R-square = 0.5117568

Change = 0.05507586

> fit9<-lm(Murder~Population+Illiteracy,data=statesDel)> shrinkage(fit9)

Original R-square = 0.6637134

10 Fold Cross-Validated R-square = 0.6444582

Change = 0.01925517

解释：

只含有Population+Illiteracy的模型比全变量模型泛化性能更好，删除了alaska和Nevada减少的最少

变量的重要性

相对权重法

relweights <- function(fit, ...) {

R <- cor(fit$model)

nvar <- ncol(R)

rxx <- R[2:nvar, 2:nvar]

rxy <- R[2:nvar, 1]

svd <- eigen(rxx)

evec <- svd$vectors

ev <- svd$values

delta <- diag(sqrt(ev))

# correlations between original predictors and new orthogonal variables

lambda <- evec %\*% delta %\*% t(evec)

lambdasq <- lambda^2

# regression coefficients of Y on orthogonal variables

beta <- solve(lambda) %\*% rxy

rsquare <- colSums(beta^2)

rawwgt <- lambdasq %\*% beta^2

import <- (rawwgt/rsquare) \* 100

lbls <- names(fit$model[2:nvar])

rownames(import) <- lbls

colnames(import) <- "Weights"

# plot results

barplot(t(import), names.arg = lbls, ylab = "% of R-Square",

xlab = "Predictor Variables", main = "Relative Importance of Predictor Variables",

sub = paste("R-Square = ", round(rsquare, digits = 3)),

...)

return(import)

}

> fit <- lm(Murder ~ Population + Illiteracy + Income + + Frost, data = states)

> relweights(fit, col = "lightgrey")

Weights

Population 14.723401

Illiteracy 59.000195

Income 5.488962

Frost 20.787442

Illiteracy 有最大的相对重要性