线性回归：机器学习基础技术

Qingming Lin

2025-01-15

Table of Contents

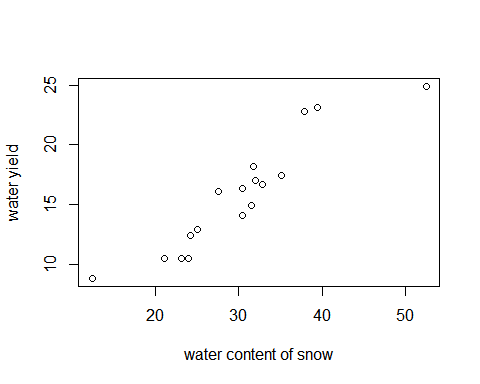
library(alr3)  
library(tidyverse)  
data(snake)  
dim(snake)

## [1] 17 2

head(snake)

## X Y  
## 1 23.1 10.5  
## 2 32.8 16.7  
## 3 31.8 18.2  
## 4 32.0 17.0  
## 5 30.4 16.3  
## 6 24.0 10.5

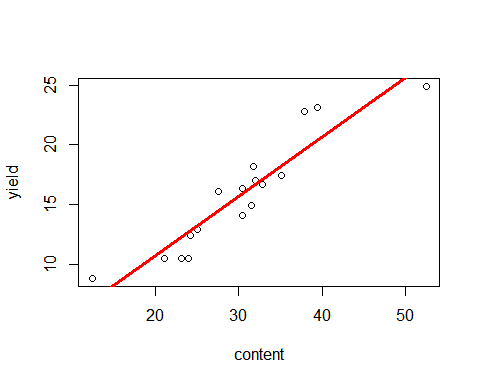
names(snake) <- c("content", "yield")  
attach(snake)  
plot(content, yield, xlab = "water content of snow", ylab = "water yield")



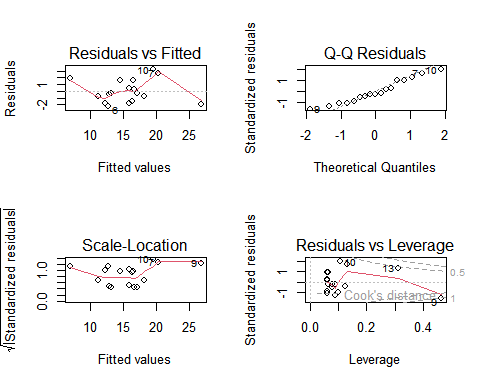
yield.fit <- lm(yield ~ content)  
summary(yield.fit)

##   
## Call:  
## lm(formula = yield ~ content)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1793 -1.5149 -0.3624 1.6276 3.1973   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.72538 1.54882 0.468 0.646   
## content 0.49808 0.04952 10.058 4.63e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.743 on 15 degrees of freedom  
## Multiple R-squared: 0.8709, Adjusted R-squared: 0.8623   
## F-statistic: 101.2 on 1 and 15 DF, p-value: 4.632e-08

plot(content, yield)  
abline(yield.fit, lwd=3, col="red")



par(mfrow = c(2,2))  
plot(yield.fit)



data(water)  
str(water)

## 'data.frame': 43 obs. of 8 variables:  
## $ Year : int 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 ...  
## $ APMAM : num 9.13 5.28 4.2 4.6 7.15 9.7 5.02 6.7 10.5 9.1 ...  
## $ APSAB : num 3.58 4.82 3.77 4.46 4.99 5.65 1.45 7.44 5.85 6.13 ...  
## $ APSLAKE: num 3.91 5.2 3.67 3.93 4.88 4.91 1.77 6.51 3.38 4.08 ...  
## $ OPBPC : num 4.1 7.55 9.52 11.14 16.34 ...  
## $ OPRC : num 7.43 11.11 12.2 15.15 20.05 ...  
## $ OPSLAKE: num 6.47 10.26 11.35 11.13 22.81 ...  
## $ BSAAM : int 54235 67567 66161 68094 107080 67594 65356 67909 92715 70024 ...

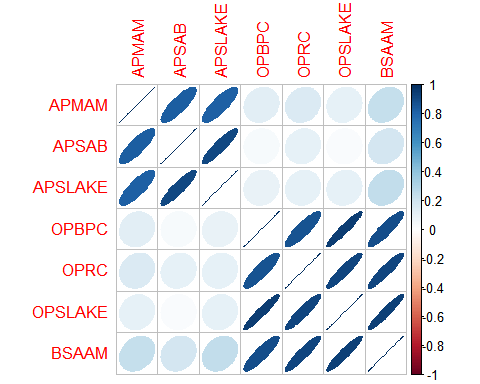
socal.water <- water[ ,-1]  
head(socal.water)

## APMAM APSAB APSLAKE OPBPC OPRC OPSLAKE BSAAM  
## 1 9.13 3.58 3.91 4.10 7.43 6.47 54235  
## 2 5.28 4.82 5.20 7.55 11.11 10.26 67567  
## 3 4.20 3.77 3.67 9.52 12.20 11.35 66161  
## 4 4.60 4.46 3.93 11.14 15.15 11.13 68094  
## 5 7.15 4.99 4.88 16.34 20.05 22.81 107080  
## 6 9.70 5.65 4.91 8.88 8.15 7.41 67594

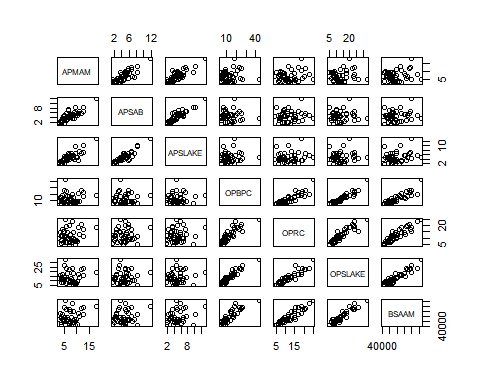
library(corrplot)  
water.cor <- cor(socal.water)  
water.cor

## APMAM APSAB APSLAKE OPBPC OPRC OPSLAKE  
## APMAM 1.0000000 0.82768637 0.81607595 0.12238567 0.1544155 0.10754212  
## APSAB 0.8276864 1.00000000 0.90030474 0.03954211 0.1056396 0.02961175  
## APSLAKE 0.8160760 0.90030474 1.00000000 0.09344773 0.1063836 0.10058669  
## OPBPC 0.1223857 0.03954211 0.09344773 1.00000000 0.8647073 0.94334741  
## OPRC 0.1544155 0.10563959 0.10638359 0.86470733 1.0000000 0.91914467  
## OPSLAKE 0.1075421 0.02961175 0.10058669 0.94334741 0.9191447 1.00000000  
## BSAAM 0.2385695 0.18329499 0.24934094 0.88574778 0.9196270 0.93843604  
## BSAAM  
## APMAM 0.2385695  
## APSAB 0.1832950  
## APSLAKE 0.2493409  
## OPBPC 0.8857478  
## OPRC 0.9196270  
## OPSLAKE 0.9384360  
## BSAAM 1.0000000

corrplot(water.cor, method = "ellipse")



pairs(~ ., data = socal.water)



library(leaps)  
fit <- lm(BSAAM ~ ., data = socal.water)  
summary(fit)

##   
## Call:  
## lm(formula = BSAAM ~ ., data = socal.water)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12690 -4936 -1424 4173 18542   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15944.67 4099.80 3.889 0.000416 \*\*\*  
## APMAM -12.77 708.89 -0.018 0.985725   
## APSAB -664.41 1522.89 -0.436 0.665237   
## APSLAKE 2270.68 1341.29 1.693 0.099112 .   
## OPBPC 69.70 461.69 0.151 0.880839   
## OPRC 1916.45 641.36 2.988 0.005031 \*\*   
## OPSLAKE 2211.58 752.69 2.938 0.005729 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7557 on 36 degrees of freedom  
## Multiple R-squared: 0.9248, Adjusted R-squared: 0.9123   
## F-statistic: 73.82 on 6 and 36 DF, p-value: < 2.2e-16

sub.fit <- regsubsets(BSAAM ~ ., data = socal.water)  
best.summary <- summary(sub.fit)  
names(best.summary)

## [1] "which" "rsq" "rss" "adjr2" "cp" "bic" "outmat" "obj"

which.min(best.summary$rss)

## [1] 6

par(mfrow = c(1,2))  
plot(best.summary$cp, xlab = "number of features", ylab = "cp")  
plot(sub.fit, scale = "Cp")

