#call up and explore the data

data(anscombe)

attach(anscombe)

anscombe

cor(x1, y1) #correlation of x1 and y1

cor(x2, y1) #correlation of x2 and y2

par(mfrow=c(2,2)) #create a 2x2 grid for plotting

plot(x1, y1, main="Plot 1")

plot(x2, y2, main="Plot 2")

plot(x3, y3, main="Plot 3")

plot(x4, y4, main="Plot 4")

install.packages("alr3")

library(alr3)

data(snake)

attach(snake)

dim(snake)

head(snake)

names(snake) = c("content", "yield")

attach(snake) #reattach data with new names

head(snake)

#produce a scatterplot

plot(content, yield, xlab="water content of snow", ylab="water yield",

main="Scatterplot of Snow vs. Yield")

#build a linear model

yield.fit = lm(yield~content)

summary(yield.fit)

plot(content, yield)

abline(yield.fit, lwd=3, col="black")

par(mfrow=c(2,2))

plot(yield.fit)

qqPlot(yield.fit)

data(water)

dim(water)

str(water)

head(water)

socal.water = water[ ,-1]

head(socal.water)

library(corrplot)

water.cor = cor(socal.water)

water.cor

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

library(leaps)

fit=lm(BSAAM~., data=socal.water)

sub.fit = regsubsets(BSAAM~., data=socal.water)

best.summary = summary(sub.fit)

names(best.summary)

which.min(best.summary$rss)

par(mfrow=c(1,2))

plot(best.summary$cp, xlab="number of features", ylab="cp")

plot(sub.fit, scale="Cp")

which.min(best.summary$bic)

which.max(best.summary$adjr2)

best.fit = lm(BSAAM~APSLAKE+OPRC+OPSLAKE, data=socal.water)

summary(best.fit)

par(mfrow=c(2,2))

plot(best.fit)

vif(best.fit)

plot(socal.water$OPRC, socal.water$OPSLAKE, xlab="OPRC", ylab="OPSLAKE")

best.summary$adjr2

fit.2 = lm(BSAAM~APSLAKE+OPSLAKE, data=socal.water)

summary(fit.2)

par(mfrow=c(2,2))

plot(fit.2)

vif(fit.2)

library(lmtest)

bptest(fit.2)

plot(fit.2$fitted.values, socal.water$BSAAM, xlab="predicted", ylab="actual", main="Predicted vs. Actual")

#ggplot2 for pretty picture

socal.water["Actual"] = water$BSAAM

socal.water["Forecast"] = NA

socal.water$Forecast = predict(fit.2)

head(socal.water)

library(ggplot2)

ggplot(socal.water, aes(x=Forecast, y=Actual))+geom\_point() + geom\_smooth(method=lm) + labs(title = "Forecast versus Actuals")

library(MPV)

PRESS(best.fit)

PRESS(fit.2)

PRESS.best = sum((resid(best.fit)/(1-hatvalues(best.fit)))^2)

PRESS.fit.2 = sum((resid(fit.2)/(1-hatvalues(fit.2)))^2)

PRESS.best

PRESS.fit.2

library(ISLR)

data(Carseats)

str(Carseats)

sales.fit = lm(Sales~Advertising+ShelveLoc, data=Carseats)

summary(sales.fit)

contrasts(Carseats$ShelveLoc)

library(MASS)

data(Boston)

str(Boston)

value.fit = lm(medv~lstat\*age, data=Boston)

summary(value.fit)