#best subsets

library(leaps)

bridge <- read.delim("C:/Teaching@cofc/Math 550/Chapter 6/bridge.txt")

logt<-log(bridge$Time,base=exp(1))

logd<-log(bridge$DArea,base=exp(1))

logc<-log(bridge$CCost,base=exp(1))

logl<-log(bridge$Length,base=exp(1))

logs<-log(bridge$Spans,base=exp(1))

logdw<- log(bridge$Dwgs,base=exp(1))

forsub <- data.frame(logt,logd,logc,logdw,logl,logs)

regfit.full = **regsubsets**(logt ~ ., data = forsub, nvmax = 19)

reg.summary = **summary**(regfit.full)

#for each case report 2 best subsets

regfit.full2= **regsubsets**(logt ~ ., data = forsub, nvmax = 19,nbest=2)

reg.summary2 = **summary**(regfit.full2)

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

**plot**(reg.summary$rss ,xlab="Number of Predictors ",ylab="RSS",type="l")

**plot**(reg.summary$adjr2 ,xlab="Number of Predictors ", ylab="Adjusted RSq",type="l")

**plot**(reg.summary$cp ,xlab="Number of Predictors ",ylab="Cp", type='l')

**plot**(reg.summary$bic ,xlab="Number of Predictors ",ylab="BIC",type='l')

lm2<-lm(logt~logdw+logs,data=forsub)

summary(lm2)

lm3<-lm(logt~logdw+logs+logc,data=forsub)

summary(lm3)

lmall<-lm(logt~logdw+logs+logc+logd+logl,data=forsub)

out<-summary(lmall)

#note BIC here did not have the other terms in the formula and it works for model comparisons

reg.summary$adjr2

reg.summary$bic

reg.summary$cp

lm2<-lm(logt~logdw+logs,data=forsub)

summary(lm2)

lm3<-lm(logt~logdw+logs+logc,data=forsub)

summary(lm3)

#AIC,BIC, AICc calculation by definition

p<-c(1,2,3,4,5)

v<-45\*log(reg.summary$rss/45,base=exp(1))

aic<-v+45+45\*log(2\*pi, base=exp(1)) +2\*(p+2)

bic<-v+45+45\*log(2\*pi, base=exp(1))+(p+2)\*log(45, base=exp(1))

aicc<-aic+2\*(p+2)\*(p+3)/(45-p-1)

cp<-reg.summary$rss/(sum(out$residual^2)/out$df[2])-(45-2\*(p+1))

reg.summary$bic-bic

# by definition using AIC, BIC functions

lm1<-lm(logt~logdw)

lm2<-lm(logt~logdw+logs)

lm3<-lm(logt~logc+ logdw+logs)

lm4<-lm(logt~logd+ logc+ logdw+logs)

lm5<- lm(logt~logd+ logc+ logdw+logs+logl)

AICV<-c(AIC(lm1),AIC(lm2),AIC(lm3),AIC(lm4),AIC(lm5))

BICV<- c(BIC(lm1),BIC(lm2),BIC(lm3),BIC(lm4),BIC(lm5))

AICCV<-c(AIC(lm1),AIC(lm2),AIC(lm3),AIC(lm4),AIC(lm5)) +2\*(p+2)\*(p+3)/(45-p-1)

# by definition using likelihood functions

A1cV<-c(-2\*logLik(lm1), -2\*logLik(lm2), -2\*logLik(lm3), -2\*logLik(lm4), -2\*logLik(lm5))+ 2\*(p+2)

B1cV<- c(-2\*logLik(lm1), -2\*logLik(lm2), -2\*logLik(lm3), -2\*logLik(lm4), -2\*logLik(lm5))+(p+2)\*log(45, base=exp(1))

A1ccV<-A1cV+ 2\*(p+2)\*(p+3)/(45-p-1)

#Backward elimination based on p-value

log<-lm(logt~logd+logc+logdw+logl+logs)

#see the details using p-value/t

summary(log)

logb1<-lm(logt~logd+logc+logdw+logs)

summary(logb1)

#remove logl

logb2<-lm(logt~ logc+logdw+logs)

summary(logb2)

#remove logd

logb4<-lm(logt~ logdw+logs)

summary(logb4)

#remove logc

# advanced fucntion

fullmodb<-lm(logt~1, data=forsub)

nullmodb<-lm(logt~ logd+logc+logdw+logl+logs,data=forsub)

reg1b<-step(nullmodb,scope=list(lower=fullmodb,upper=nullmodb), direction ="backward")

# understand what is going on

logwl<-lm(logt~logd+logc+logdw+logs)

#reported AIC is not including the other terms but good for model comparison purpose

AICReporteddifference<-AIC(logwl)-AIC(log)

AICReported<-AIC(logwl)

AICStarted<-AIC(log)

RSSReported<-sum(logwl$residuals^2)

RSSReportedall<-sum(log$residuals^2)

#final model from backward based on p-value

best1<-lm(logt~logdw+logs)

summary(best1)

#final model from backward based on AIC

best2<-lm(logt~ logs+logc+logdw)

summary(best2)

nullmod<-lm(logt~1,data=forsub)

fullmod<-lm(logt~ logd+logc+logdw+logl+logs,data=forsub)

reg1A<-step(nullmod,scope=list(lower=nullmod, upper=fullmod), direction="forward")

reg1A

#explain what is going on in forward selection based on p-value

lm11<-lm(logt~logd)

lm12<-lm(logt~logc)

lm13<-lm(logt~logdw)

lm14<-lm(logt~logl)

lm15<-lm(logt~logs)

summary(lm11)

summary(lm12)

summary(lm13)

summary(lm14)

summary(lm15)

#logdw enter the model first

lm21<-lm(logt~logdw+logd)

lm22<-lm(logt~logdw+logc)

lm23<-lm(logt~logdw+logl)

lm24<-lm(logt~logdw+logs)

summary(lm21)

summary(lm22)

summary(lm23)

summary(lm24)

#logs enter the model

lm31<-lm(logt~logdw+logs+logd)

lm32<-lm(logt~logdw+logs+logc)

lm33<-lm(logt~logdw+logs+logl)

summary(lm31)

summary(lm32)

summary(lm33)

#no more useful predictor to add in the model. Final model with predictors logdw and logs

fullmodb<-lm(logt~1, data=forsub)

nullmodb<-lm(logt~ logd+logc+logdw+logl+logs,data=forsub)

reg1c<-step(fullmodb,scope=list(lower=fullmodb,upper=nullmodb), direction ="both",k=2)

# k = 2 gives the AIC: k = log(n) is referred to as BIC

reg1c<-step(fullmodb,scope=list(lower=fullmodb,upper=nullmodb), direction ="both",k=log(nrow(forsub)))

# For predictive ability

prostateTraining <- read.delim("C:/Teaching@cofc/Math 550/Chapter 7/prostateTraining.txt")

library(tidyverse)

library(GGally)

pairs(~lpsa+lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=prostateTraining)

lm<-lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=prostateTraining)

res<-rstandard(lm)

plot(res~lcavol,data= prostateTraining)

plot(res~lweight,data= prostateTraining)

plot(res~age,data= prostateTraining)

plot(res~lbph,data= prostateTraining)

plot(res~svi,data= prostateTraining)

plot(res~lcp,data= prostateTraining)

plot(res~gleason,data= prostateTraining)

plot(res~pgg45,data= prostateTraining)

plot(res~fitted(lm))

plot(lpsa~fitted(lm),data=prostateTraining)

plot(lm)

avPlots(lm, layout=c(2,2))

mmps(lm,layout=c(2,2), ylab=" lpsa ")

summary(lm)

library(car)

vif(lm)

library(leaps)

attach(prostateTraining)

forsub <- data.frame(lpsa,lcavol,lweight,age,lbph,svi,lcp,gleason,pgg45)

p<-c(1,2,3,4,5,6,7,8)

regfit.full = **regsubsets**(lpsa ~ ., data = forsub, nvmax = 16)

reg.summary = **summary**(regfit.full)

reg.summary$adjr2

reg.summary$bic

**par**(mfrow=**c**(1,1))

**plot**(reg.summary$adjr2 ,xlab="Number of Variables ", ylab="Adjusted RSq",type="l")

**plot**(reg.summary$bic ,xlab="Number of Variables ",ylab="BIC",type='l')

n<-67

v<-n\*log(reg.summary$rss/n,base=exp(1))

aic<-v+2\*(p+2)+n+n\*log(2\*pi, base=exp(1))

bic<-v+n+n\*log(2\*pi, base=exp(1))+(p+2)\*log(n, base=exp(1))

aicc<-aic+2\*(p+2)\*(p+3)/(n-p-1)

lm2<-lm(lpsa~lcavol+lweight)

summary(lm2)

lm4<-lm(lpsa~ lcavol+lweight+ lbph+ svi)

summary(lm4)

lm7<-lm(lpsa~ lcavol+lweight+age+ lbph+ svi+lcp+pgg45)

summary(lm7)

prostateTest <- read.delim("C:/Teaching@cofc/Math 550/Chapter 7/prostateTest.txt")

attach(prostateTest)

lm2<-lm(lpsa~lcavol+lweight)

summary(lm2)

lm4<-lm(lpsa~ lcavol+lweight+ lbph+ svi)

summary(lm4)

lm7<-lm(lpsa~ lcavol+lweight+age+ lbph+ svi+lcp+pgg45)

summary(lm7)

# exclude 45th observation  
newdata <- subset(forsub, original\_case!=69)

p<-c(1,2,3,4,5,6,7,8)

regfit.fullw = **regsubsets**(lpsa ~ ., data = newdata, nvmax = 16)

reg.summaryw = **summary**(regfit.fullw)

reg.summaryw$adjr2

reg.summaryw$bic

reg.summaryw$aic

n<-66

v<-n\*log(reg.summaryw$rss/n,base=exp(1))

aic<-v+2\*(p+2)+n+n\*log(2\*pi, base=exp(1))

bic<-v+n+n\*log(2\*pi, base=exp(1))+(p+2)\*log(n, base=exp(1))

aicc<-aic+2\*(p+2)\*(p+3)/(n-p-1)

reg.summary

reg.summaryw

# note case 45 has dramatic effect on variable selection

plot(forsub$lweight,forsub$lpsa, pch=3 , col = "blue")

lmtraining<-lm(lpsa~lweight, data=forsub)

abline(lmtraining,lty=1, , col = "blue")

points(prostateTest$lweight,prostateTest$lpsa, pch=2 , col = "red")

lmtest<-lm(lpsa~lweight, data= prostateTest)

abline(lmtest,lty=2, col = "red")

# case 45 does not standout but case 9 calls for further investigation

lmt<-lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=prostateTest)

avPlots(lmt)

avPlots(lm)

# case 45 in the training data and case 9 in the test data need further investigation. If they can not be simply removed, then other variable selection techniques needed.

#exercise

null<-lm(Y~1,data=Mantel)

full<-lm(Y~ X1+X2+X3,data=Mantel)

ex1A<-step(null,scope=list(lower=null, upper=full), direction="forward")

ex1B<-step(null,scope=list(lower=null, upper=full), direction="forward",k=log(5))