basic steps of linear regression

ruo

Thursday, April 09, 2015

Step 1: read the data into R studio

setwd=("C:/Users/Ruo/Documents/R")  
fueleff<-read.csv("C:/Users/Ruo/Documents/R/FuelEfficiency.csv")

step2 :learn the basic info about the dataset. There is highly corralition between some virables, which means we need to drop some of them.

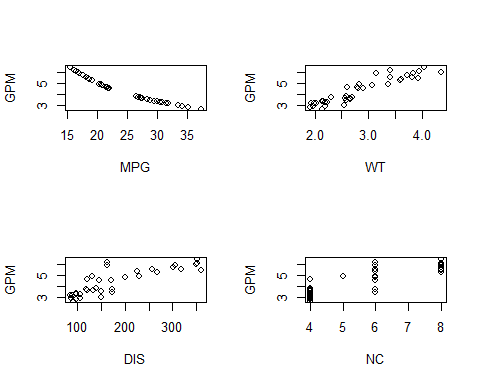
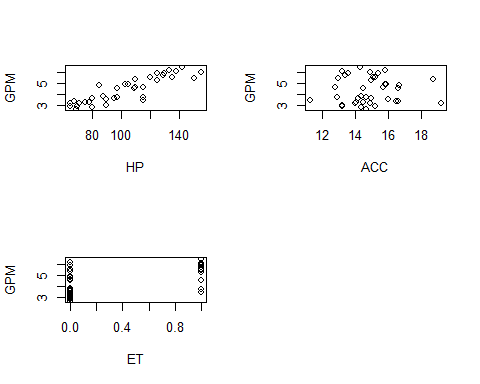
head(fueleff)

## MPG GPM WT DIS NC HP ACC ET  
## 1 16.9 5.917 4.360 350 8 155 14.9 1  
## 2 15.5 6.452 4.054 351 8 142 14.3 1  
## 3 19.2 5.208 3.605 267 8 125 15.0 1  
## 4 18.5 5.405 3.940 360 8 150 13.0 1  
## 5 30.0 3.333 2.155 98 4 68 16.5 0  
## 6 27.5 3.636 2.560 134 4 95 14.2 0

cor(fueleff)

## MPG GPM WT DIS NC HP  
## MPG 1.00000000 -0.98079724 -0.90307083 -0.7860481 -0.8055110 -0.8712821  
## GPM -0.98079724 1.00000000 0.92626656 0.8229098 0.8411880 0.8876992  
## WT -0.90307083 0.92626656 1.00000000 0.9507647 0.9166777 0.9172204  
## DIS -0.78604807 0.82290984 0.95076469 1.0000000 0.9402812 0.8717993  
## NC -0.80551105 0.84118805 0.91667774 0.9402812 1.0000000 0.8638473  
## HP -0.87128209 0.88769915 0.91722045 0.8717993 0.8638473 1.0000000  
## ACC -0.05677359 0.03307093 -0.03357386 -0.1434174 -0.1292436 -0.2526211  
## ET -0.49816677 0.52061208 0.66736606 0.7746636 0.8311721 0.7202350  
## ACC ET  
## MPG -0.05677359 -0.4981668  
## GPM 0.03307093 0.5206121  
## WT -0.03357386 0.6673661  
## DIS -0.14341745 0.7746636  
## NC -0.12924363 0.8311721  
## HP -0.25262113 0.7202350  
## ACC 1.00000000 -0.3102336  
## ET -0.31023357 1.0000000

from the graphs, we could know there is some variables showing linear regression relationship to the observation. We can try it!

step3: build model initally

model.full<-lm(GPM~.,data=fueleff)  
summary(model.full)

##   
## Call:  
## lm(formula = GPM ~ ., data = fueleff)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33622 -0.13499 -0.00486 0.08701 0.48325   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.9864020 1.1309678 4.409 0.000123 \*\*\*  
## MPG -0.1151678 0.0160110 -7.193 5.27e-08 \*\*\*  
## WT 0.0860549 0.2949136 0.292 0.772450   
## DIS -0.0003056 0.0017785 -0.172 0.864705   
## NC 0.2152198 0.0819830 2.625 0.013497 \*   
## HP 0.0076517 0.0047066 1.626 0.114471   
## ACC 0.0136109 0.0282887 0.481 0.633905   
## ET -0.4770695 0.1774648 -2.688 0.011608 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1928 on 30 degrees of freedom  
## Multiple R-squared: 0.9775, Adjusted R-squared: 0.9722   
## F-statistic: 185.8 on 7 and 30 DF, p-value: < 2.2e-16

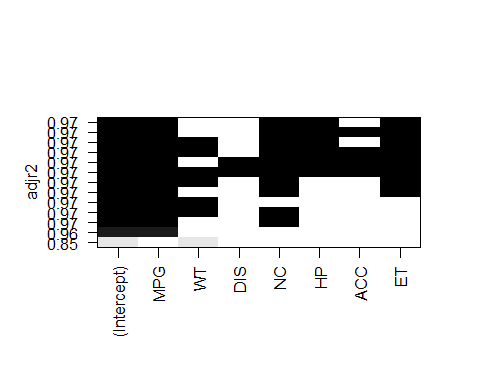
step4:select variables 1)all-subsets regression

library(leaps)

## Warning: package 'leaps' was built under R version 3.2.1

leaps=regsubsets(GPM~.,data=fueleff,nbest=2,nvmax=7)# 2 best model for each number of predictors  
out=summary(leaps)  
tab=cbind(out$which,out$rsq,out$adjr2,out$cp)  
tab

## (Intercept) MPG WT DIS NC HP ACC ET   
## 1 1 1 0 0 0 0 0 0 0.9619632 0.9609066 16.612808  
## 1 1 0 1 0 0 0 0 0 0.8579697 0.8540244 154.989469  
## 2 1 1 1 0 0 0 0 0 0.9708716 0.9692071 6.759095  
## 2 1 1 0 0 1 0 0 0 0.9694125 0.9676646 8.700641  
## 3 1 1 0 0 1 0 0 1 0.9735647 0.9712322 5.175543  
## 3 1 1 1 0 1 0 0 0 0.9714533 0.9689345 7.984976  
## 4 1 1 0 0 1 1 0 1 0.9770590 0.9742783 2.525946  
## 4 1 1 1 0 1 0 0 1 0.9750609 0.9720379 5.184709  
## 5 1 1 0 0 1 1 1 1 0.9773851 0.9738516 4.091979  
## 5 1 1 1 0 1 1 0 1 0.9772136 0.9736532 4.320247  
## 6 1 1 1 0 1 1 1 1 0.9774321 0.9730641 6.029534  
## 6 1 1 0 1 1 1 1 1 0.9773903 0.9730142 6.085146  
## 7 1 1 1 1 1 1 1 1 0.9774543 0.9721936 8.000000

The resulting table shows the trade-off between model size and model fit  we got MPG+NC+HP+ET

2)stepwise method.forward stepwise and backward stepwise

library(MASS)  
full.model<-lm(GPM~.,data=fueleff)  
reduced.model<-step(full.model,direction="backward")

## Start: AIC=-118.1  
## GPM ~ MPG + WT + DIS + NC + HP + ACC + ET  
##   
## Df Sum of Sq RSS AIC  
## - DIS 1 0.00110 1.1159 -120.062  
## - WT 1 0.00316 1.1179 -119.992  
## - ACC 1 0.00860 1.1234 -119.808  
## <none> 1.1148 -118.100  
## - HP 1 0.09821 1.2130 -116.891  
## - NC 1 0.25608 1.3708 -112.242  
## - ET 1 0.26853 1.3833 -111.898  
## - MPG 1 1.92259 3.0373 -82.011  
##   
## Step: AIC=-120.06  
## GPM ~ MPG + WT + NC + HP + ACC + ET  
##   
## Df Sum of Sq RSS AIC  
## - WT 1 0.00232 1.1182 -121.984  
## - ACC 1 0.01080 1.1267 -121.696  
## <none> 1.1159 -120.062  
## - HP 1 0.10536 1.2212 -118.634  
## - NC 1 0.26597 1.3818 -113.939  
## - ET 1 0.29000 1.4059 -113.284  
## - MPG 1 2.24381 3.3597 -80.178  
##   
## Step: AIC=-121.98  
## GPM ~ MPG + NC + HP + ACC + ET  
##   
## Df Sum of Sq RSS AIC  
## - ACC 1 0.01613 1.1343 -123.439  
## <none> 1.1182 -121.984  
## - HP 1 0.17739 1.2956 -118.388  
## - ET 1 0.34748 1.4657 -113.700  
## - NC 1 0.50051 1.6187 -109.927  
## - MPG 1 2.27647 3.3947 -81.785  
##   
## Step: AIC=-123.44  
## GPM ~ MPG + NC + HP + ET  
##   
## Df Sum of Sq RSS AIC  
## <none> 1.1343 -123.44  
## - HP 1 0.17277 1.3071 -120.05  
## - ET 1 0.35093 1.4852 -115.20  
## - NC 1 0.51432 1.6486 -111.23  
## - MPG 1 3.08956 4.2239 -75.48

min.model <- lm(GPM~1,data=fueleff)  
forward.model<-step(min.model,direction="forward",scope=(~MPG + WT + DIS + NC + HP + ACC + ET))

## Start: AIC=12  
## GPM ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + MPG 1 47.564 1.881 -110.226  
## + WT 1 42.422 7.023 -60.161  
## + HP 1 38.963 10.482 -44.942  
## + NC 1 34.987 14.458 -32.722  
## + DIS 1 33.483 15.962 -28.961  
## + ET 1 13.401 36.043 1.991  
## <none> 49.445 12.004  
## + ACC 1 0.054 49.390 13.963  
##   
## Step: AIC=-110.23  
## GPM ~ MPG  
##   
## Df Sum of Sq RSS AIC  
## + WT 1 0.44047 1.4402 -118.36  
## + NC 1 0.36832 1.5124 -116.51  
## + DIS 1 0.34929 1.5314 -116.03  
## + HP 1 0.22556 1.6551 -113.08  
## <none> 1.8807 -110.23  
## + ET 1 0.06739 1.8133 -109.61  
## + ACC 1 0.02536 1.8554 -108.74  
##   
## Step: AIC=-118.37  
## GPM ~ MPG + WT  
##   
## Df Sum of Sq RSS AIC  
## <none> 1.4402 -118.36  
## + NC 1 0.0287653 1.4115 -117.13  
## + ET 1 0.0248612 1.4154 -117.03  
## + HP 1 0.0067145 1.4335 -116.54  
## + ACC 1 0.0008125 1.4394 -116.39  
## + DIS 1 0.0007110 1.4395 -116.38

the best AIC is -123.44 GPM ~ MPG + NC + HP + ET

step5: build model Then we try lm with MPG + NC + HP + ET

model.2<-lm(GPM~MPG + NC + HP + ET,data=fueleff)  
summary(model.2)

##   
## Call:  
## lm(formula = GPM ~ MPG + NC + HP + ET, data = fueleff)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36789 -0.11118 0.00162 0.08547 0.48024   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.523792 0.718593 7.687 7.50e-09 \*\*\*  
## MPG -0.120947 0.012757 -9.481 6.05e-11 \*\*\*  
## NC 0.226629 0.058588 3.868 0.000488 \*\*\*  
## HP 0.007115 0.003174 2.242 0.031804 \*   
## ET -0.500716 0.156707 -3.195 0.003071 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1854 on 33 degrees of freedom  
## Multiple R-squared: 0.9771, Adjusted R-squared: 0.9743   
## F-statistic: 351.4 on 4 and 33 DF, p-value: < 2.2e-16

Adjusted R-squared: 0.9743 ,even better then the ARS of the model with all variables, which is 0.9722

step6:model validation

par(mfrow = c(2,2))   
plot(model.2,which=c(1:4))

