Covariates Selection Analysis

SIT Wai Tang

```
rm(list=ls())
library(aod)
library(car)
```

```
## Loading required package: carData
```

```
library(data.table)
library(leaps)
```

The motor trend car road test data comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

mpg: Miles/(US) gallon

cyl: Number of cylinders

disp: Displacement (cu.in.)

hp: Gross horsepower

drat: Rear axle ratio

wt: Weight (1000 lbs)

qsec: 1/4 mile time

vs: Engine (0 = V-shaped, 1 = straight)

am: Transmission (0 = automatic, 1 = manual)

gear: Number of forward gears

carb: Number of carburetors

```
data(mtcars)
mtcars$vs = factor(mtcars$vs)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec       vs          am        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   0:18   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1:14   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71          Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85          Mean   :0.4062  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90          3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90          Max.   :1.0000  
##       gear            carb      
##  Min.   :3.000   Min.   :1.000  
##  1st Qu.:3.000   1st Qu.:2.000  
##  Median :4.000   Median :2.000  
##  Mean   :3.688   Mean   :2.812  
##  3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :5.000   Max.   :8.000  
```

Filter out the best regression model to Predict $\boldsymbol{mpg}$ from all 10 variables of full model, Based on $Backward$ selection and the adjusted $\boldsymbol{R^2}$ criterion.

full model: $\boldsymbol{E(mpg) = \beta 0 + \beta 1 cyl + \beta 2 hp + \beta 3 wt + \beta 4 qsec + \beta 5 vs + \beta 6 disp + \beta 7 drat + \beta 8 am + \beta 9 gear + \beta 10 carb}$

```
library(leaps)
regfit_full = regsubsets(mpg~., data=mtcars, method="backward")
summary(regfit_full)
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ ., data = mtcars, method = "backward")
## 10 Variables  (and intercept)
##      Forced in Forced out
## cyl      FALSE      FALSE
## disp     FALSE      FALSE
## hp       FALSE      FALSE
## drat     FALSE      FALSE
## wt       FALSE      FALSE
## qsec     FALSE      FALSE
## vs1      FALSE      FALSE
## am       FALSE      FALSE
## gear     FALSE      FALSE
## carb     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##          cyl disp hp  drat wt  qsec vs1 am  gear carb
## 1  ( 1 ) " " " "  " " " "  "*" " "  " " " " " "  " "
## 2  ( 1 ) " " " "  " " " "  "*" "*"  " " " " " "  " "
## 3  ( 1 ) " " " "  " " " "  "*" "*"  " " "*" " "  " "
## 4  ( 1 ) " " " "  "*" " "  "*" "*"  " " "*" " "  " "
## 5  ( 1 ) " " "*"  "*" " "  "*" "*"  " " "*" " "  " "
## 6  ( 1 ) " " "*"  "*" "*"  "*" "*"  " " "*" " "  " "
## 7  ( 1 ) " " "*"  "*" "*"  "*" "*"  " " "*" "*"  " "
## 8  ( 1 ) " " "*"  "*" "*"  "*" "*"  " " "*" "*"  "*"
```

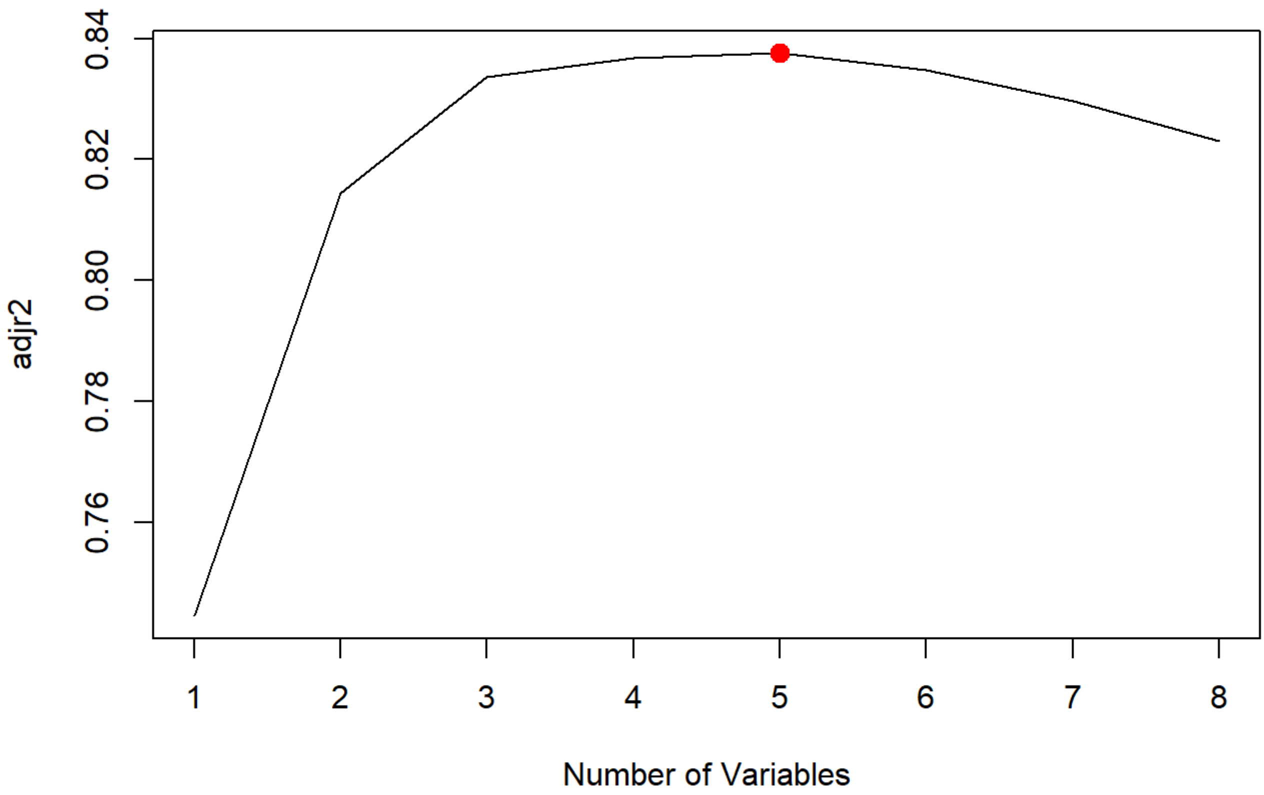
```
summary(regfit_full)$adjr2 ; which.max(summary(regfit_full)$adjr2)
```

```
## [1] 0.7445939 0.8144448 0.8335561 0.8367919 0.8375334 0.8347177 0.8296261
## [8] 0.8230390
```

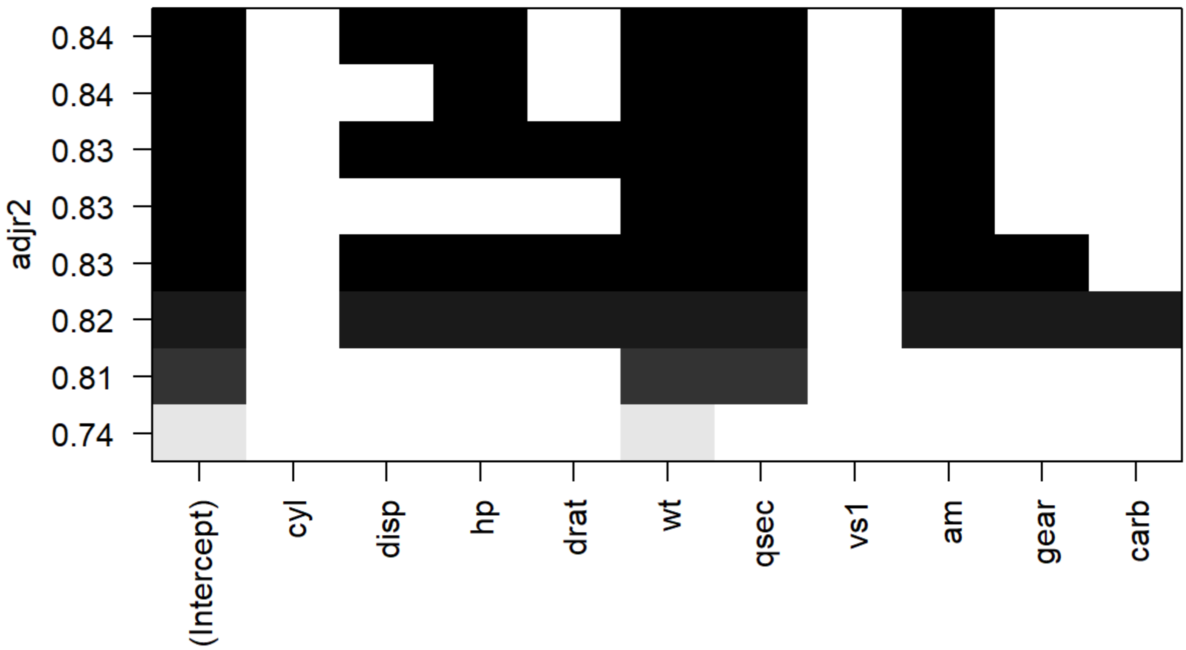
```
## [1] 5
```

Visualization

```
plot(summary(regfit_full)$adjr2, xlab = "Number of Variables", ylab = "adjr2", type = "l")
adjr2_max = which.max(summary(regfit_full)$adjr2) #
points(adjr2_max, summary(regfit_full)$adjr2[adjr2_max], col = "red", cex = 2, pch = 20)
```




```
plot(regfit_full, scale="adjr2")
```




```
fit.reduced = lm(mpg~disp+hp+wt+qsec+am, data=mtcars)
summary(fit.reduced)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5399 -1.7398 -0.3196  1.1676  4.5534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.36190    9.74079   1.474  0.15238
## disp         0.01124    0.01060   1.060  0.29897
## hp          -0.02117    0.01450  -1.460  0.15639
## wt          -4.08433    1.19410  -3.420  0.00208 **
## qsec         1.00690    0.47543   2.118  0.04391 *
## am           3.47045    1.48578   2.336  0.02749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.429 on 26 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8375
## F-statistic: 32.96 on 5 and 26 DF,  p-value: 1.844e-10
```

```
shapiro.test(fit.reduced$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.reduced$residuals
## W = 0.95389, p-value = 0.1858
```

The test statistics W is Large, reject the null hypothesis that the random sample is normally distributed.

the estimated values, standard errors and p-values for$\boldsymbol{\beta 2}$,$\boldsymbol{\beta 3}$,$\boldsymbol{\beta 4}$,$\boldsymbol{\beta 6}$, $\boldsymbol{\beta 8}$.

```
list("estimated values"=coef(summary.lm(fit.reduced))[1:6,1],"standard errors"=coef(summary.lm(fit.redu
ced))[1:6,2])
```

```
## $`estimated values`
## (Intercept)        disp          hp          wt        qsec          am
## 14.36190396  0.01123765 -0.02117055 -4.08433206  1.00689683  3.47045340
##
## $`standard errors`
## (Intercept)        disp          hp          wt        qsec          am
##  9.74079485  0.01060333  0.01450469  1.19409972  0.47543287  1.48578009
```

Compare Full model with reduced model

```
fit.full = lm(mpg~., data=mtcars)
anova(fit.reduced, fit.full)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ disp + hp + wt + qsec + am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 153.44
## 2     21 147.49  5    5.9434 0.1692 0.9711
```

The p-value is 0.9711, which is > 0.05, hence we do not reject the null hypothesis and conclude that the reduced model is better.

Base on backward selection with 5 variables and adjusted R^2 criterion ,The best regression model to predict mpg is: $E(mpg) = \beta 0 + \beta 2hp + \beta 3wt + \beta 4qsec + \beta 6disp + \beta 8am$

Filter out the best regression model to Predict mpg from all 10 variables of full model, Based on $Stepwise$ selection and the BIC criterion.

```
fit.regstep = regsubsets(mpg~., data=mtcars, method="seqrep")
summary(fit.regstep)
```

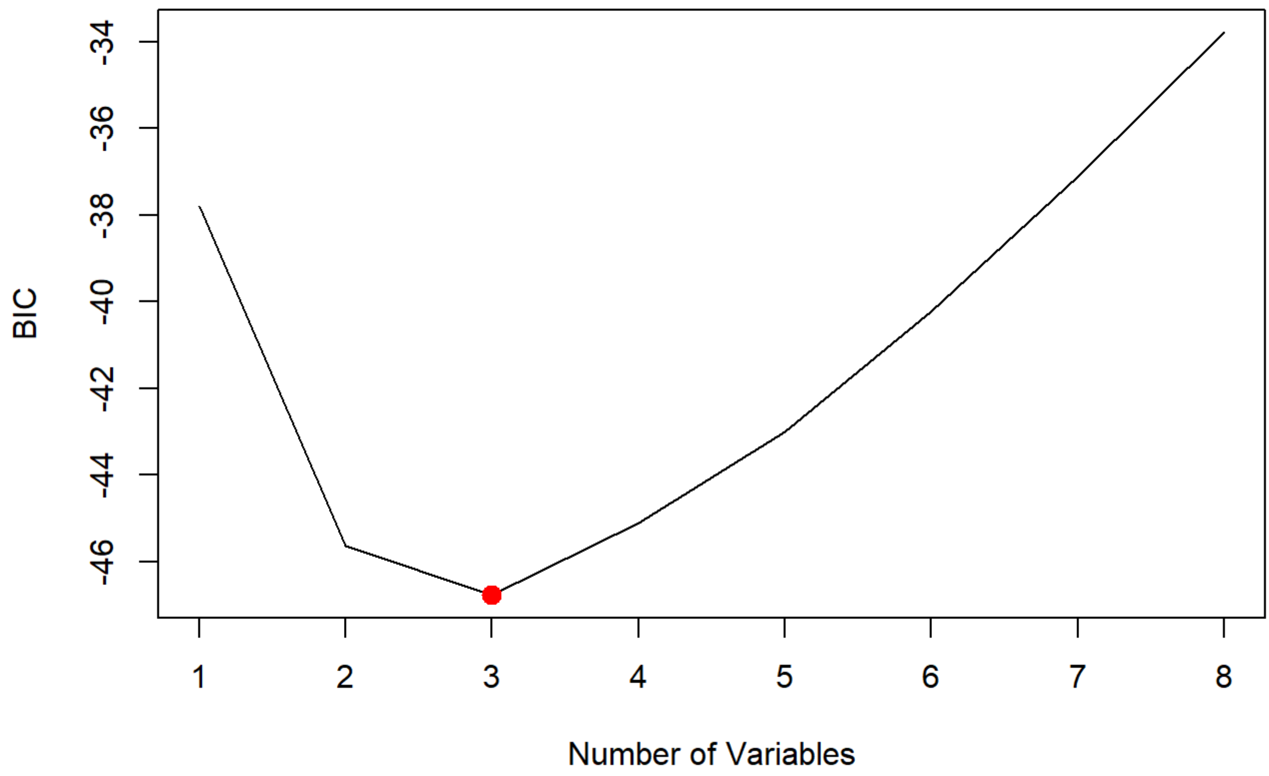
```
## Subset selection object
## Call: regsubsets.formula(mpg ~ ., data = mtcars, method = "seqrep")
## 10 Variables  (and intercept)
##      Forced in Forced out
## cyl      FALSE      FALSE
## disp     FALSE      FALSE
## hp       FALSE      FALSE
## drat     FALSE      FALSE
## wt       FALSE      FALSE
## qsec     FALSE      FALSE
## vs1      FALSE      FALSE
## am       FALSE      FALSE
## gear     FALSE      FALSE
## carb     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: 'sequential replacement'
##          cyl disp hp  drat wt  qsec vs1 am  gear carb
## 1  ( 1 ) " " " "  " " " "  "*" " "  " " " " " "  " "
## 2  ( 1 ) "*" "*"  " " " "  " " " "  " " " " " "  " "
## 3  ( 1 ) "*" " "  "*" " "  "*" " "  " " " " " "  " "
## 4  ( 1 ) " " " "  "*" " "  "*" "*"  " " "*" " "  " "
## 5  ( 1 ) " " "*"  "*" " "  "*" "*"  " " "*" " "  " "
## 6  ( 1 ) " " "*"  "*" "*"  "*" "*"  " " "*" " "  " "
## 7  ( 1 ) " " "*"  "*" "*"  "*" "*"  " " "*" "*"  " "
## 8  ( 1 ) "*" "*"  "*" "*"  "*" "*"  "*" "*" " "  " "
```

```
summary(fit.regstep)$bic ; which.min(summary(fit.regstep)$bic)
```

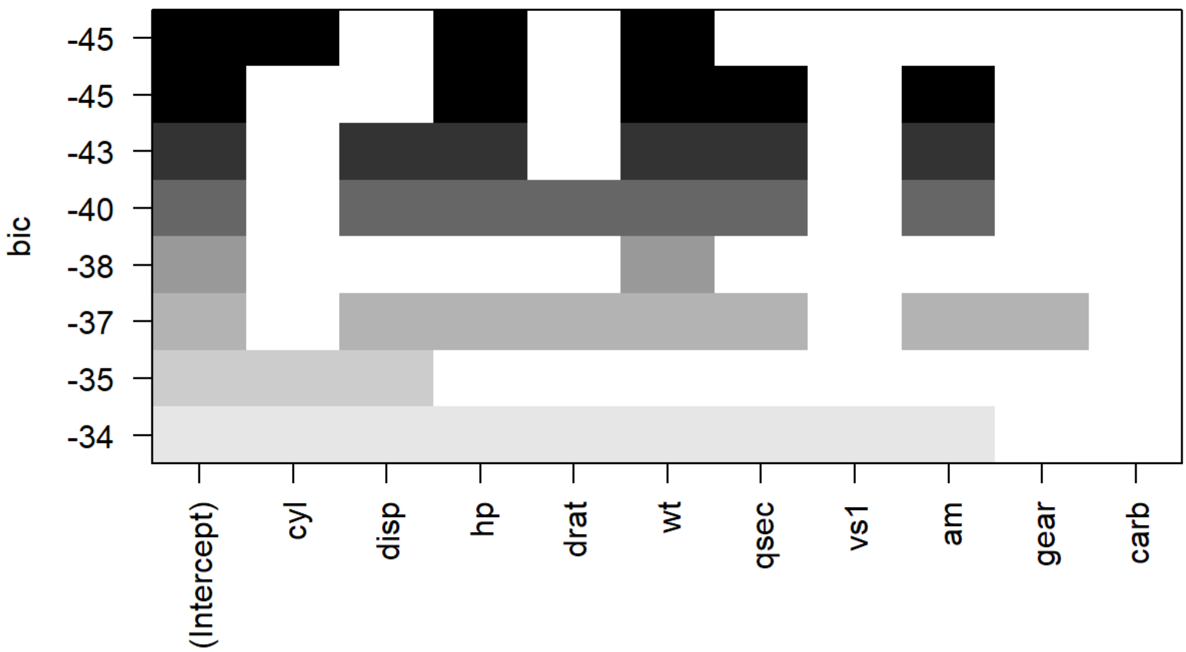
```
## [1] -37.79462 -35.21267 -45.41594 -45.09947 -42.98713 -40.22663 -37.09630
## [8] -33.55642
```

```
## [1] 3
```

```
plot(summary(regfit_full)$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min = which.min(summary(regfit_full)$bic) #
points(bic_min, summary(regfit_full)$bic[bic_min], col = "red", cex = 2, pch = 20)
```




```
plot(fit.regstep, scale="bic")
```



by using Stepwise selection with 3 variables and adjusted BIC criterion ,The best model to predict mpg is: $\boldsymbol{E(mpg)}$ = $\boldsymbol{\beta 0}$ +$\boldsymbol{\beta 1 cyl}$ + $\boldsymbol{\beta 2 hp}$ + $\boldsymbol{\beta 3 wt}$

R^2;Backward: $\boldsymbol{E(mpg)}$ = $\boldsymbol{\beta 0}$ + $\boldsymbol{\beta 2 hp}$ + $\boldsymbol{\beta 3 wt}$ + $\boldsymbol{\beta 4 qsec}$ + $\boldsymbol{\beta 6 disp}$ + $\boldsymbol{\beta 8 am}$

BIC;Stepwise: $\boldsymbol{E(mpg)}$ = $\boldsymbol{\beta 0}$ +$\boldsymbol{\beta 1 cyl}$ + $\boldsymbol{\beta 2 hp}$ + $\boldsymbol{\beta 3 wt}$

Test significance of Variables vs, drat, disp and crab at $\boldsymbol{\alpha}$ = $\boldsymbol{0.05}$ level.

$H_0 : \beta_5 = \beta_7 = \beta_9 = \beta_{10} = 0 \quad vs \quad H_1$: at least two coefficients are different

```
anova(lm(mpg~.-vs-drat-gear-carb, data=mtcars), fit.full)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ (cyl + disp + hp + drat + wt + qsec + vs + am + gear +
##     carb) - vs - drat - gear - carb
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     25 150.99
## 2     21 147.49  4    3.4967 0.1245  0.972
```

p-value = 0.972 > 0.05, not reject H0, the variables vs, drat, disp and crab are not significant.

Variable Selection Analysis, base on criterion: BIC, Mallow’s $\boldsymbol{Cp}$ and adjusted $\boldsymbol{R2}$

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
# (1) cyl, hp, wt
fit1 = lm(mpg~cyl+ hp+ wt, data=mtcars) # +cyl?

# (2) hp, wt, qsec
fit2 = lm(mpg~hp+ wt+ qsec, data=mtcars) # +qsec?

# (3) hp, wt ,disp
fit3 = lm(mpg~hp+ wt+ disp, data=mtcars) # +disp?

# (4) hp, wt, am
fit4 = lm(mpg~hp+ wt+ am, data=mtcars) # +am?

# (5) cyl, hp, wt , am
fit5 = lm(mpg~cyl+ hp+ wt+ am, data=mtcars) # +cyl+am?
```

```
VSA = cbind(c(BIC(fit1),BIC(fit2),
              BIC(fit3),BIC(fit4),BIC(fit5)),
            c(ols_mallows_cp(fit1, fit.full),
              ols_mallows_cp(fit2, fit.full),
              ols_mallows_cp(fit3, fit.full),
              ols_mallows_cp(fit4, fit.full),
              ols_mallows_cp(fit5, fit.full)),
            c(summary(fit1)$adj.r.squared,summary(fit2)$adj.r.squared,
              summary(fit3)$adj.r.squared,summary(fit4)$adj.r.squared,
              summary(fit5)$adj.r.squared))
colnames(VSA) = c("BIC", "Mallow’s C_P","adjusted R_2")
VSA
```

```
##           BIC Mallow’s C_P adjusted R_2
## [1,] 162.8053     1.146922    0.8263446
## [2,] 164.4713     2.490799    0.8170643
## [3,] 165.9717     3.762433    0.8082829
## [4,] 163.4635     1.669529    0.8227357
## [5,] 165.0481     2.203986    0.8266657
```

According to the BIC, model 1 is the best with the smallest value 162.8053

According to the Mallow’s Cp, model 1 is the best with the smallest value 1.146922

According to adjusted R square, model 5 is the best with the largest value 0.8266657, model 1 is 2nd best with the largest value 0.8263446.

Therefore, $\boldsymbol{E(mpg)} = \boldsymbol{\beta 0} + \boldsymbol{\beta 1 cyl} + \boldsymbol{\beta 2 hp} + \boldsymbol{\beta 3 wt}$ is preferred.

Predict the $\boldsymbol{mpg}$ value and 95% confidence interval when Test value: (cyl=4, disp=110, hp=93,drat=3.85, wt=2.5, qsec=16.3,vs= 1,am= 1, gear=3, carb=1)

```
fit_preferred = lm(mpg~cyl+hp+wt, data=mtcars)
fit_preferred
```

```
## 
## Call:
## lm(formula = mpg ~ cyl + hp + wt, data = mtcars)
## 
## Coefficients:
## (Intercept)          cyl           hp           wt  
##    38.75179     -0.94162     -0.01804     -3.16697
```

```r
predict(fit_preferred, newdata=data.frame(cyl=4, hp=93, wt=2.5), se.fit=T, interval="confidence")
```

```
## $fit
##        fit      lwr      upr
## 1 25.39034 23.85084 26.92985
##
## $se.fit
## [1] 0.7515618
##
## $df
## [1] 28
##
## $residual.scale
## [1] 2.511548
```

The predicted value of mpg is 25.39034, and 95% confidence interval is (23.85084, 26.92985)