Stat 415/615, Lab 5. Model (variable) selection in MLR

Jun Lu

Stat 415/615 Regression, 2023

Contents

Comments and explanations are not included here. We'll discuss them in class.

Consider the **Surgical Unit Example** from text, p.350, and refer to Handout6_ModelSelection.pdf

```
sudata<-read.table("../DataSets/SurgicalUnit.txt", header=T)
head(sudata, 2)
```

```
##   bloodclotting prognostic enzyme liver age gender alchoholM alcoholeS survival
## 1           6.7         62     81  2.59  50      0         1         0      695
## 2           5.1         59     66  1.70  39      0         0         0      403
##    lnY
## 1 6.54
## 2 6.00
```

```
tail(sudata, 2)
```

```
##    bloodclotting prognostic enzyme liver age gender alchoholM alcoholeS
## 53           6.4         59     85  2.33  63      0         1         0
## 54           8.8         78     72  3.20  56      0         0         0
##    survival  lnY
## 53      550 6.31
## 54      651 6.48
```

We will skip the descriptive summary for now, and focus on the variable selection.

1 Compute model selection criteria

- You can copy and paste the following code to define a function `selcri()` that computes R^2, adjusted-R^2, AIC, SBC (aka. BIC), and PRESS.

```
selcri<-function(lmout)
{
 n <- length(lmout$fit)
 rsq <- summary(lmout)$r.sq
 adj.rsq <- summary(lmout)$adj.r.sq
 aic <- extractAIC(lmout)[2]
```

```
 bic <- extractAIC(lmout, k = log(n))[2]
 press <- sum((lmout$residuals/(1 - hatvalues(lmout)))^2)
 cbind(rsq, adj.rsq, aic, bic, press)
}
```

```
sureg1<-lm(lnY ~ bloodclotting, data=sudata)
sureg2<-lm(lnY ~ bloodclotting+ prognostic+ enzyme+ liver , data=sudata)

selcri(sureg1)
```

```
##             rsq    adj.rsq       aic       bic    press
## [1,] 0.06062567 0.04256078 -77.10872 -73.13076 13.50244
```

```
selcri(sureg2)
```

```
##            rsq  adj.rsq       aic       bic    press
## [1,] 0.7580574 0.738307 -144.3605 -134.4155 4.085347
```

- Mellows-C requires careful consideration of "all available predictors." We'll assume "all available predictors" refers to `bloodclotting`, `prognostic`, `enzyme` and `liver` in this example.

```
anova(sureg1)
```

```
## Analysis of Variance Table
##
## Response: lnY
##               Df  Sum Sq Mean Sq F value  Pr(>F)
## bloodclotting  1  0.7761 0.77606   3.356 0.07269 .
## Residuals     52 12.0248 0.23125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(sureg2)
```

```
## Analysis of Variance Table
##
## Response: lnY
##               Df Sum Sq Mean Sq F value    Pr(>F)
## bloodclotting  1 0.7761  0.7761  12.278 0.0009893 ***
## prognostic     1 2.5890  2.5890  40.961 5.640e-08 ***
## enzyme         1 6.3149  6.3149  99.911 2.043e-13 ***
## liver          1 0.0238  0.0238   0.377 0.5420460
## Residuals     49 3.0971  0.0632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mallow1<-12.025/0.063 - (length(sureg1$fit)-2*2)
mallow1
```

```
## [1] 140.873
```

2 Stepwise selection based on AIC (or BIC)

```
# help("step")
step(sureg2, direction="both") #Stepwise selection by AIC
```

```
## Start:  AIC=-144.36
```

```
## lnY ~ bloodclotting + prognostic + enzyme + liver
##
##                 Df Sum of Sq    RSS     AIC
## - liver          1    0.0238 3.1209 -145.95
## <none>                       3.0971 -144.36
## - bloodclotting  1    0.5317 3.6288 -137.81
## - prognostic     1    1.8870 4.9840 -120.67
## - enzyme         1    3.4807 6.5778 -105.69
##
## Step:  AIC=-145.95
## lnY ~ bloodclotting + prognostic + enzyme
##
##                 Df Sum of Sq    RSS      AIC
## <none>                       3.1209 -145.947
## + liver          1    0.0238 3.0971 -144.360
## - bloodclotting  1    1.2025 4.3234 -130.347
## - prognostic     1    2.6725 5.7934 -114.543
## - enzyme         1    6.3149 9.4358  -88.201
```

```
##
## Call:
## lm(formula = lnY ~ bloodclotting + prognostic + enzyme, data = sudata)
##
## Coefficients:
##   (Intercept)  bloodclotting     prognostic         enzyme
##       3.76888        0.09540        0.01334        0.01643
```

Remarks:

- To select variables by BIC (aka SBC), set k = log(sample size).

```
step(sureg2, direction="both", k=log(length(sureg2$fit)))
```

```
## Start:  AIC=-134.42
## lnY ~ bloodclotting + prognostic + enzyme + liver
##
##                 Df Sum of Sq    RSS      AIC
## - liver          1    0.0238 3.1209 -137.991
## <none>                       3.0971 -134.416
## - bloodclotting  1    0.5317 3.6288 -129.849
## - prognostic     1    1.8870 4.9840 -112.712
## - enzyme         1    3.4807 6.5778  -97.729
##
## Step:  AIC=-137.99
## lnY ~ bloodclotting + prognostic + enzyme
##
##                 Df Sum of Sq    RSS      AIC
## <none>                       3.1209 -137.991
## + liver          1    0.0238 3.0971 -134.416
## - bloodclotting  1    1.2025 4.3234 -124.380
## - prognostic     1    2.6725 5.7934 -108.576
## - enzyme         1    6.3149 9.4358  -82.234
```

```
##
## Call:
## lm(formula = lnY ~ bloodclotting + prognostic + enzyme, data = sudata)
##
```

```
## Coefficients:
##   (Intercept)  bloodclotting     prognostic         enzyme
##       3.76888        0.09540        0.01334        0.01643
```

- If you want to keep certain predictors in the model regardless how they affect the AIC or BIC, use the "scope" argument inside the `step()` function. Refer to the help file for `step()` function for more details.

3 Best Subset selection (we need to install a package: leaps)

Run the following line if package "leaps" has not been downloaded and installed.

```
install.packages("leaps")
```

```
library(leaps)
# help(regsubsets)

temp<-regsubsets(lnY ~ bloodclotting+ prognostic+ enzyme+ liver + age + gender +
    alchoholM + alcoholeS, data=sudata)
bestsub<-summary(temp)
bestsub
```

```
## Subset selection object
## Call: regsubsets.formula(lnY ~ bloodclotting + prognostic + enzyme +
##     liver + age + gender + alchoholM + alcoholeS, data = sudata)
## 8 Variables  (and intercept)
##               Forced in Forced out
## bloodclotting     FALSE      FALSE
## prognostic        FALSE      FALSE
## enzyme            FALSE      FALSE
## liver             FALSE      FALSE
## age               FALSE      FALSE
## gender            FALSE      FALSE
## alchoholM         FALSE      FALSE
## alcoholeS         FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          bloodclotting prognostic enzyme liver age gender alchoholM alcoholeS
## 1  ( 1 ) " "           " "        "*"    " "   " " " "    " "       " "
## 2  ( 1 ) " "           "*"        "*"    " "   " " " "    " "       " "
## 3  ( 1 ) " "           "*"        "*"    " "   " " " "    " "       "*"
## 4  ( 1 ) "*"           "*"        "*"    " "   " " " "    " "       "*"
## 5  ( 1 ) "*"           "*"        "*"    " "   " " "*"    " "       "*"
## 6  ( 1 ) "*"           "*"        "*"    " "   "*" "*"    " "       "*"
## 7  ( 1 ) "*"           "*"        "*"    " "   "*" "*"    "*"       "*"
## 8  ( 1 ) "*"           "*"        "*"    "*"   "*" "*"    "*"       "*"
```

```
bestsub$adjr2
```

```
## [1] 0.4153921 0.6490153 0.7632496 0.8145138 0.8192299 0.8221892 0.8213983
## [8] 0.8175210
```

```
bestsub$rsq
```

```
## [1] 0.4264224 0.6622600 0.7766506 0.8285128 0.8362837 0.8423187 0.8449872
## [8] 0.8450650
```

```
bestsub$bic
```

```
## [1] -22.03858 -46.64891 -64.99103 -75.27040 -73.78559 -71.82480 -68.75751
## [8] -64.79561
```

```
bestsub$cp
```

```
## [1] 116.592349  50.094657  18.870569   5.807484   5.550465   5.797630   7.022579
## [8]   9.000000
```

—— This is the end of Lab 5. ——