CH3_Regression_ex2

Philip oh

변수 선택

```
library(datasets)
library(MASS)
library(ISLR)
head(Credit)
```

```
##   ID  Income Limit Rating Cards Age Education Gender Student Married
## 1  1  14.891  3606    283     2  34        11   Male      No     Yes
## 2  2 106.025  6645    483     3  82        15 Female     Yes     Yes
## 3  3 104.593  7075    514     4  71        11   Male      No      No
## 4  4 148.924  9504    681     3  36        11 Female      No      No
## 5  5  55.882  4897    357     2  68        16   Male      No     Yes
## 6  6  80.180  8047    569     4  77        10   Male      No      No
##   Ethnicity Balance
## 1 Caucasian     333
## 2     Asian     903
## 3     Asian     580
## 4     Asian     964
## 5 Caucasian     331
## 6 Caucasian    1151
```

```
credit.fit = lm(Balance ~ ., data = Credit)
summary(credit.fit)
```

```
##
## Call:
## lm(formula = Balance ~ ., data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -166.48  -77.62  -14.37   56.21  316.52
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -487.07424   36.73407 -13.259  < 2e-16 ***
## ID                   0.04105    0.04343   0.945   0.3452
## Income              -7.80740    0.23431 -33.321  < 2e-16 ***
## Limit                0.19052    0.03279   5.811  1.3e-08 ***
## Rating               1.14249    0.49100   2.327   0.0205 *
## Cards               17.83639    4.34324   4.107  4.9e-05 ***
## Age                 -0.62955    0.29449  -2.138   0.0332 *
## Education           -1.09831    1.59817  -0.687   0.4924
## GenderFemale        -9.54615    9.98431  -0.956   0.3396
## StudentYes         426.16715   16.73077  25.472  < 2e-16 ***
## MarriedYes          -8.78055   10.36758  -0.847   0.3976
## EthnicityAsian      16.85752   14.12112   1.194   0.2333
## EthnicityCaucasian   9.29289   12.24194   0.759   0.4483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.8 on 387 degrees of freedom
## Multiple R-squared:  0.9552, Adjusted R-squared:  0.9538
## F-statistic: 687.7 on 12 and 387 DF,  p-value: < 2.2e-16
```

단계적 선택방법(Stepwise selection)

```
aic.credit = stepAIC(credit.fit, direction="both", trace = 0)
aic.credit
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Student, data = Credit)
##
## Coefficients:
## (Intercept)       Income        Limit       Rating        Cards
##   -493.7342      -7.7951       0.1937       1.0912      18.2119
##         Age   StudentYes
##     -0.6241     425.6099
```

- `direction` 에는 foward와 backward도 있다.
- 최종 모형으로 Balance ~ Income + Limit + Rating + Cards + Age + Student가 나왔다.
- `trace = 0` 으로 하면 과정을 모두 볼 필요가 없다.

```
credit.step = lm(Balance~Income+Limit+Rating+Cards+Age+Student, data=Credit)
summary(credit.step)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Student, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.00  -77.85  -11.84   56.87  313.52
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -493.73419   24.82476 -19.889  < 2e-16 ***
## Income        -7.79508    0.23342 -33.395  < 2e-16 ***
## Limit          0.19369    0.03238   5.981 4.98e-09 ***
## Rating         1.09119    0.48480   2.251   0.0250 *
## Cards         18.21190    4.31865   4.217 3.08e-05 ***
## Age           -0.62406    0.29182  -2.139   0.0331 *
## StudentYes   425.60994   16.50956  25.780  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.61 on 393 degrees of freedom
## Multiple R-squared:  0.9547, Adjusted R-squared:  0.954
## F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

- 단계적 선택방법으로 얻은 최적의 모형으로 다중회귀분석을 다시 돌려보았다.

All subset search

```
library(leaps)
```

```
all.sub = regsubsets(Balance ~ ., data = Credit, nbest=2)
summary(all.sub)
```

```
## Subset selection object
## Call: regsubsets.formula(Balance ~ ., data = Credit, nbest = 2)
## 12 Variables  (and intercept)
##                   Forced in Forced out
## ID                    FALSE      FALSE
## Income                FALSE      FALSE
## Limit                 FALSE      FALSE
## Rating                FALSE      FALSE
## Cards                 FALSE      FALSE
## Age                   FALSE      FALSE
## Education             FALSE      FALSE
## GenderFemale          FALSE      FALSE
## StudentYes            FALSE      FALSE
## MarriedYes            FALSE      FALSE
## EthnicityAsian        FALSE      FALSE
## EthnicityCaucasian    FALSE      FALSE
## 2 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          ID  Income Limit Rating Cards Age Education GenderFemale
## 1  ( 1 ) " " " "    " "   "*"    " "   " " " "       " "
## 1  ( 2 ) " " " "    "*"   " "    " "   " " " "       " "
## 2  ( 1 ) " " "*"    " "   "*"    " "   " " " "       " "
## 2  ( 2 ) " " "*"    "*"   " "    " "   " " " "       " "
## 3  ( 1 ) " " "*"    " "   "*"    " "   " " " "       " "
## 3  ( 2 ) " " "*"    "*"   " "    " "   " " " "       " "
## 4  ( 1 ) " " "*"    "*"   " "    "*"   " " " "       " "
## 4  ( 2 ) " " "*"    "*"   "*"    " "   " " " "       " "
## 5  ( 1 ) " " "*"    "*"   "*"    "*"   " " " "       " "
## 5  ( 2 ) " " "*"    "*"   " "    "*"   "*" " "       " "
## 6  ( 1 ) " " "*"    "*"   "*"    "*"   "*" " "       " "
## 6  ( 2 ) " " "*"    "*"   "*"    "*"   " " " "       "*"
## 7  ( 1 ) " " "*"    "*"   "*"    "*"   "*" " "       "*"
## 7  ( 2 ) "*" "*"    "*"   "*"    "*"   "*" " "       " "
## 8  ( 1 ) "*" "*"    "*"   "*"    "*"   "*" " "       "*"
## 8  ( 2 ) "*" "*"    "*"   "*"    "*"   "*" " "       " "
##          StudentYes MarriedYes EthnicityAsian EthnicityCaucasian
## 1  ( 1 ) " "        " "        " "            " "
## 1  ( 2 ) " "        " "        " "            " "
## 2  ( 1 ) " "        " "        " "            " "
## 2  ( 2 ) " "        " "        " "            " "
## 3  ( 1 ) "*"        " "        " "            " "
## 3  ( 2 ) "*"        " "        " "            " "
## 4  ( 1 ) "*"        " "        " "            " "
## 4  ( 2 ) "*"        " "        " "            " "
## 5  ( 1 ) "*"        " "        " "            " "
## 5  ( 2 ) "*"        " "        " "            " "
## 6  ( 1 ) "*"        " "        " "            " "
## 6  ( 2 ) "*"        " "        " "            " "
## 7  ( 1 ) "*"        " "        " "            " "
## 7  ( 2 ) "*"        " "        " "            " "
## 8  ( 1 ) "*"        " "        " "            " "
## 8  ( 2 ) "*"        " "        "*"            " "
```

- 옵션 중 nbest 를 2로 입력했기 때문에 각 변수개수별로 최적의 모형 2개씩을 구해주었다.
- 변수의 개수가 n개일 때 어떤 변수를 포함해야 최적의 모형이 되는지 알려준다. 즉, 결과창의 맨 왼쪽의 숫자가 2이면, 2개의 변수만 넣어야 할 때, 어떤 변수를 넣어야 최적의 모형이 되는지를 설명해준다.

BIC와 수정결정계수 확인하기

```
summary(all.sub)$bic
```

```
##  [1]  -535.9468  -530.7458  -814.1798  -801.5344 -1173.3585 -1164.9522
##  [7] -1198.0527 -1186.2300 -1197.0957 -1196.6003 -1195.7321 -1192.2803
## [13] -1190.8790 -1190.8732 -1185.7841 -1185.5683
```

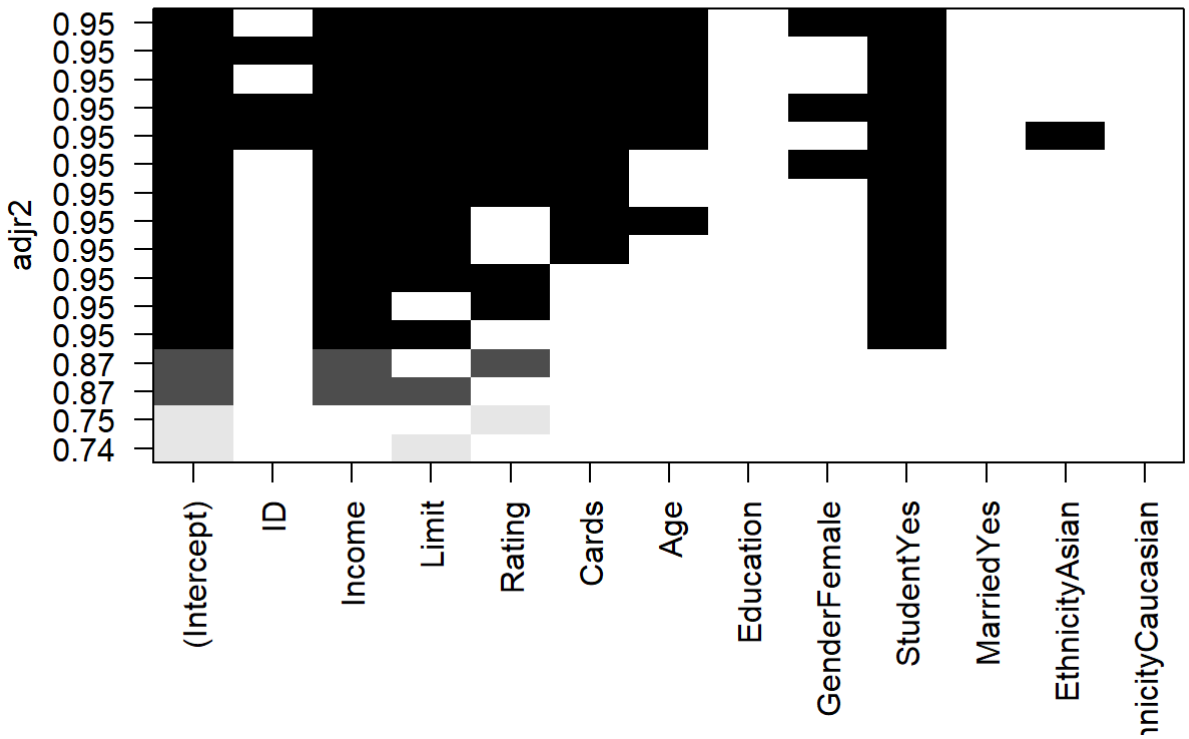
```
summary(all.sub)$adjr2
```

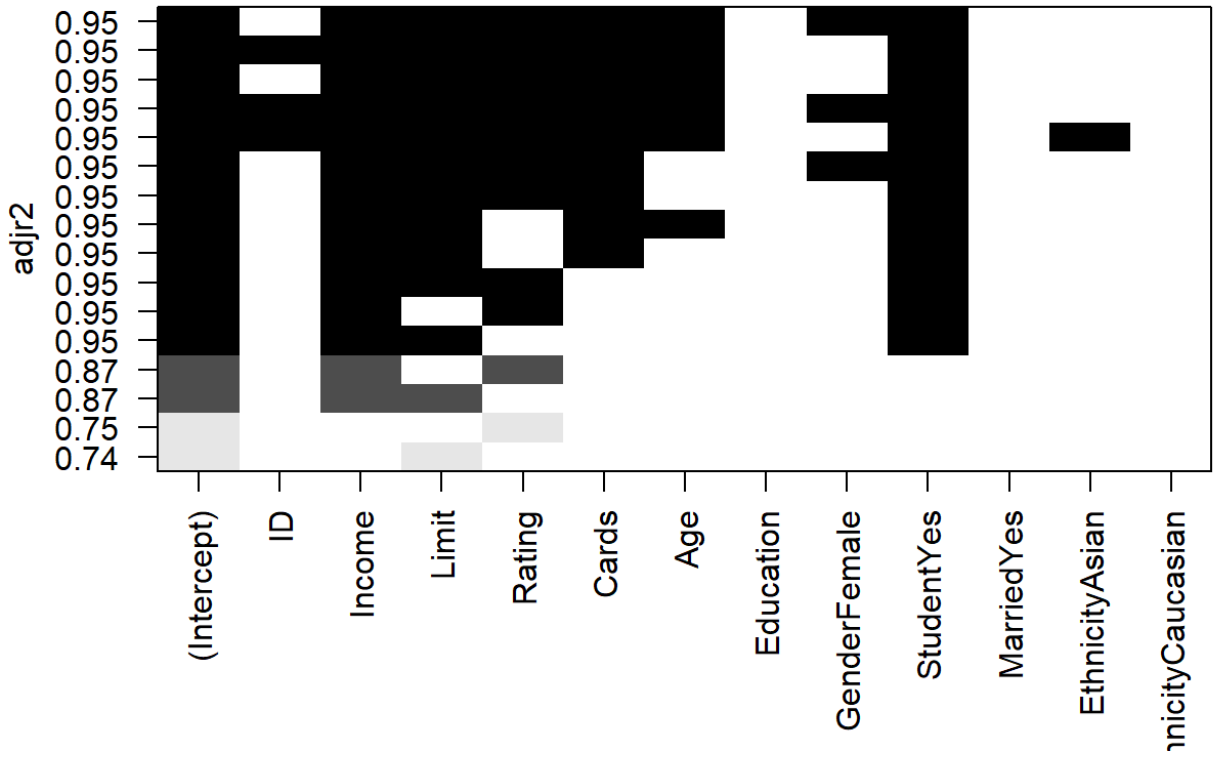
```
##  [1] 0.7452098 0.7418753 0.8744888 0.8704576 0.9494991 0.9484265 0.9531099
##  [8] 0.9517033 0.9535789 0.9535213 0.9539961 0.9535974 0.9540098 0.9540091
## [15] 0.9539954 0.9539706
```

- 위와 같은 방법으로 bic와 수정 결정 계수도 확인할 수 있다.

수정결정계수를 그림으로 확인하기

```
plot(all.sub, scale="adjr2")
```

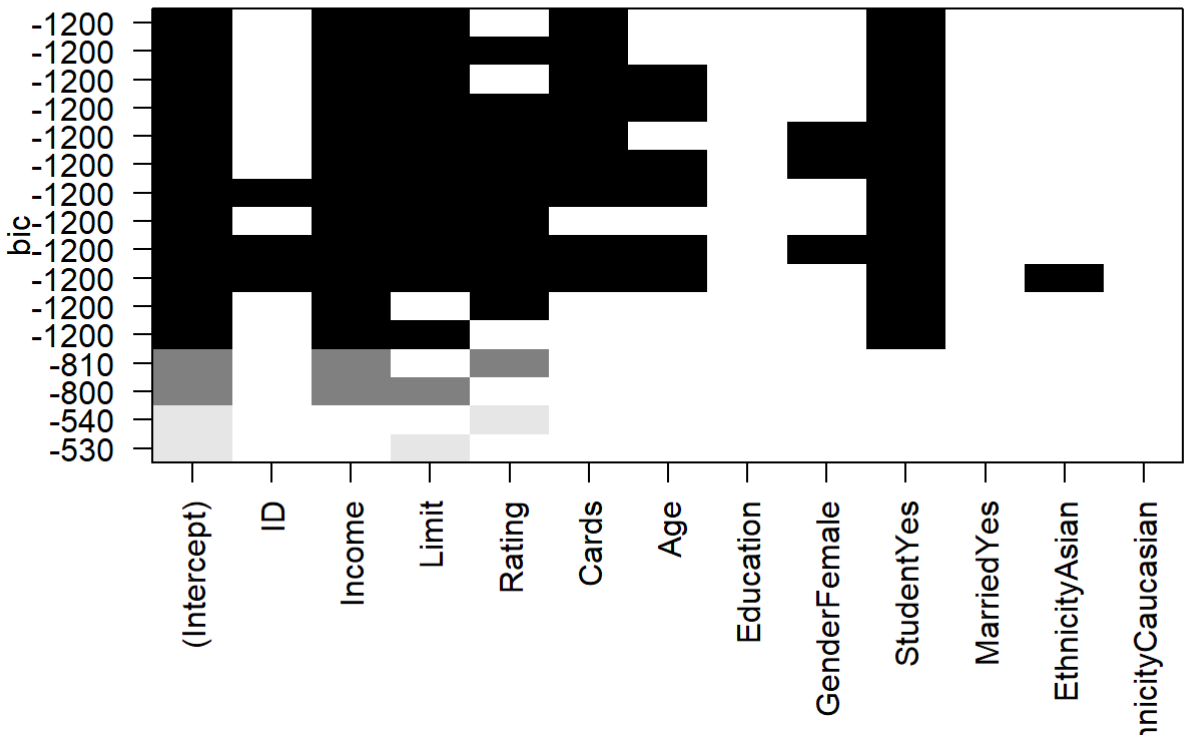


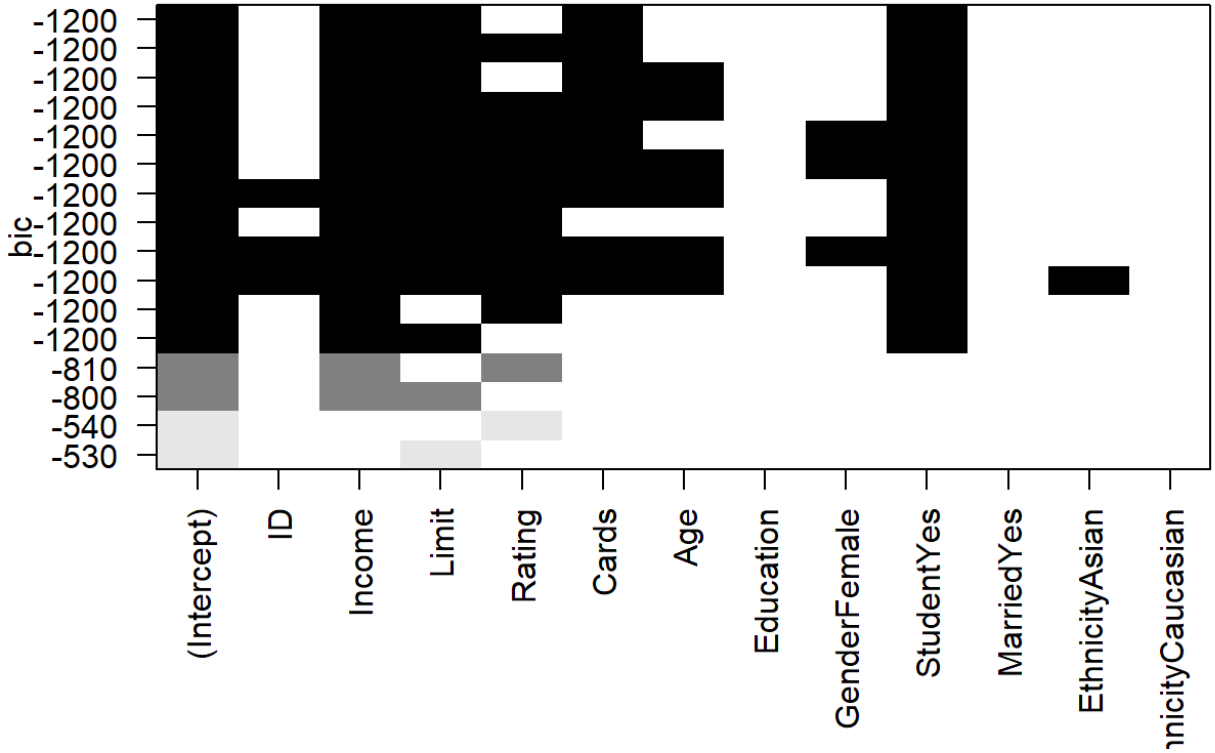


- 좀 더 쉽게 보고싶다면 `plot` 에 `scale` 옵션을 부여하면 된다.
- 수정결정계수가 높을 수록 좋은 모형이므로 위 그림에서 가장 위에 있는 변수 조합이 가장 좋은 모형이다.

BIC를 그림으로 확인하기

```
plot(all.sub)
```





- 수정결정계수와 달리 BIC는 낮을 수록 좋은 모형이다. 맨 위의 변수 조합이 가장 좋은 모형을 만든다.