STAT 500: HW7

Jasmine Mou

11/14/2017

1. Using the `teengamb` dataset with `gamble` as the response and the other variables as predictors. Implement the following variable selection methods to determine the "best" model:

```
data(teengamb, package="faraway")
lm_gamble <- lm(gamble~., data=teengamb)
summary(lm_gamble)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  22.55565063 17.1968034  1.3116188 1.967736e-01
## sex         -22.11833009  8.2111145 -2.6937062 1.011184e-02
## status        0.05223384  0.2811115  0.1858118 8.534869e-01
## income        4.96197922  1.0253923  4.8391032 1.791882e-05
## verbal       -2.95949350  2.1721503 -1.3624718 1.803109e-01
```

(a) Backward elimination
Set the α_{crit} to be 0.05. With the full model, we can see the variable `status` has the largest p-value over 0.05, and is not that significant in influencing `gamble`. Refit the model without `status`.

```
lm_gamble <- update(lm_gamble, . ~ . - status)
summary(lm_gamble)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  24.138972 14.7685884  1.634481 1.094591e-01
## sex         -22.960220  6.7705747 -3.391177 1.502436e-03
## income        4.898090  0.9551179  5.128256 6.643750e-06
## verbal       -2.746817  1.8252807 -1.504874 1.396672e-01
```

Now `verbal` becomes the predictor with the largest p-value over 0.05. Refit the model with the removal of `verbal`.

```
lm_gamble <- update(lm_gamble, . ~ . - verbal)
summary(lm_gamble)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)   4.040829  6.3943499  0.6319374 5.306977e-01
## sex         -21.634391  6.8087973 -3.1774174 2.717320e-03
## income        5.171584  0.9510477  5.4377755 2.244878e-06
```

Up to this stage, all variable's p-value are less than α_{crit} except for the intercept. Thus the best model selected with backward elimination is:

$$gamble = 4.041 - 21.634 * sex + 5.172 * income$$

(b) AIC
For each size of model p, do exhaustive search to find the variables that produce the minimum RSS.

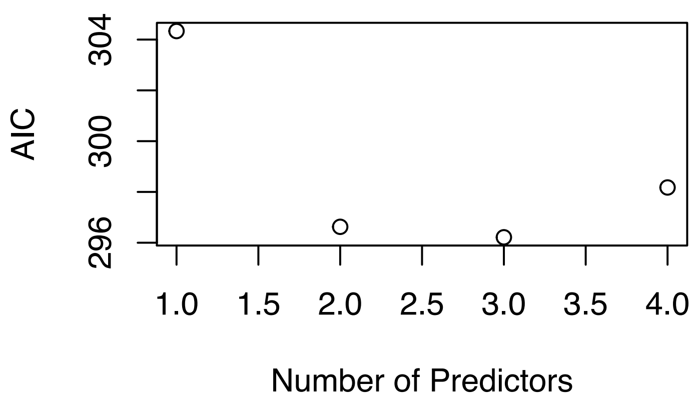
```
require(leaps)
b <- regsubsets(gamble~., data=teengamb)
rs <- summary(b)
rs$which
```

```
##   (Intercept)   sex status income verbal
## 1        TRUE FALSE  FALSE   TRUE  FALSE
## 2        TRUE  TRUE  FALSE   TRUE  FALSE
## 3        TRUE  TRUE  FALSE   TRUE   TRUE
## 4        TRUE  TRUE   TRUE   TRUE   TRUE
```

Compute and plot AIC. We can see that AIC is minimized by choosing 3 predictors, which are `income`*,* `sex`*, and* `verbal` *from the logical matrix above. Fit the linear model with these predictors. According to the fitted summary coefficients, the best model determined by AIC will be*

$$gamble = 24.139 + 4.898 * income - 22.960 * sex - 2.747 * verbal$$

```r
n = dim(teengamb)[1]
p = dim(teengamb)[2]-1
AIC <- n * log(rs$rss/n) + (2:(p+1))*2
plot(AIC ~ I(1:p), ylab="AIC", xlab="Number of Predictors")
```




```r
lm_gamble <- lm(gamble~income+sex+verbal, data=teengamb)
summary(lm_gamble)$coefficients
```

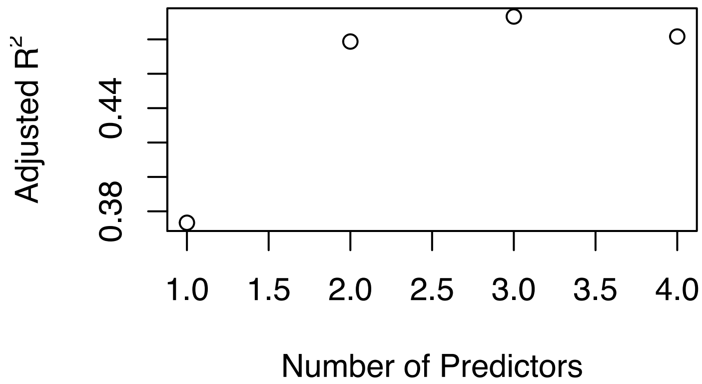
```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  24.138972 14.7685884  1.634481 1.094591e-01
## income        4.898090  0.9551179  5.128256 6.643750e-06
## sex         -22.960220  6.7705747 -3.391177 1.502436e-03
## verbal       -2.746817  1.8252807 -1.504874 1.396672e-01
```

(c) Adjusted R^2

Plot R^2 with the number of predictors used. We can see that R^2 achieves the maximum when 3 predictors are used. Thus the best model selected with Adjusted R^2 is the same model selected with AIC:

$$gamble = 24.139 + 4.898 * income - 22.960 * sex - 2.747 * verbal$$

```r
plot(rs$adjr2 ~ I(1:p), xlab="Number of Predictors", ylab=expression(paste("Adjusted ", R^2)))
```




```
which.max(rs$adjr2)
```

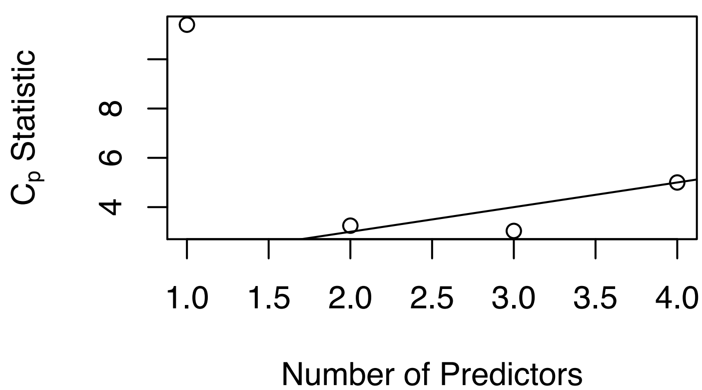
```
## [1] 3
```

(d) Mallows C_p

Plot C_p against the number of predictors used. We can see only the models with 3 and 4 predictors are on or below the C_p = p+1 line, C_p Statistic is minimized when the number of predictors is 3. Thus the best model selected with Mallows C_p is the same model selected with AIC:

$$gamble = 24.139 + 4.898 * income - 22.960 * sex - 2.747 * verbal$$

```
plot(rs$cp ~ I(1:p), xlab="Number of Predictors", ylab=expression(paste(C[p], " Statistic")))
abline(1,1)
```




```
which.min(rs$cp)
```

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

````
---
title: 'STAT 500: HW7'
author: "Jasmine Mou"
date: "11/14/2017"
output: pdf_document
---

1. Using the `teengamb` dataset with `gamble` as the response and the
other variables as predictors. Implement the following variable
selection methods to determine the “best” model:

```{r}
data(teengamb, package="faraway")
lm_gamble <- lm(gamble~., data=teengamb)
summary(lm_gamble)$coefficients
```

(a) Backward elimination
*Set the $\alpha_{crit}$ to be 0.05. With the full model, we can see the
variable `status` has the largest p-value over 0.05, and is not that
significant in influencing `gamble`. Refit the model without `status`.*
```{r}
lm_gamble <- update(lm_gamble, . ~ . - status)
summary(lm_gamble)$coefficients
```
*Now `verbal` becomes the predictor with the largest p-value over 0.05.
Refit the model with the removal of `verbal`.*
```{r}
lm_gamble <- update(lm_gamble, . ~ . - verbal)
summary(lm_gamble)$coefficients
```
*Up to this stage, all variable's p-value are less than $\alpha_{crit}$
except for the intercept. Thus the best model selected with backward
elimination is: $$ gamble = 4.041 -21.634 * sex + 5.172 * income $$ *

(b) AIC
*For each size of model p, do exhaustive search to find the variables
that produce the minimum RSS.*
```{r, warning=FALSE, message=FALSE}
require(leaps)
b <- regsubsets(gamble~., data=teengamb)
rs <- summary(b)
rs$which
```
*Compute and plot AIC. We can see that AIC is minimized by choosing 3
predictors, which are `income`, `sex`, and `verbal` from the logical
matrix above. Fit the linear model with these predictors. According to
the fitted summary coefficients, the best model determined by AIC will
be $$ gamble = 24.139 + 4.898 * income - 22.960 * sex - 2.747 * verbal
$$ *
````

```{r, fig.width=4, fig.height=3}
n = dim(teengamb)[1]
p = dim(teengamb)[2]-1
AIC <- n * log(rs$rss/n) + (2:(p+1))*2
plot(AIC ~ I(1:p), ylab="AIC", xlab="Number of Predictors")

lm_gamble <- lm(gamble~income+sex+verbal, data=teengamb)
summary(lm_gamble)$coefficients
```

(c) Adjusted R^2

*Plot R^2 with the number of predictors used. We can see that R^2 achieves the maximum when 3 predictors are used. Thus the best model selected with Adjusted R^2 is the same model selected with AIC: $$ gamble = 24.139 + 4.898 * income - 22.960 * sex - 2.747 * verbal $$*

```{r, fig.width=4, fig.height=3}
plot(rs$adjr2 ~ I(1:p), xlab="Number of Predictors",
ylab=expression(paste("Adjusted ", R^2)))
which.max(rs$adjr2)
```

(d) Mallows C_p

*Plot C_p against the number of predictors used. We can see only the models with 3 and 4 predictors are on or below the C_p = p+1 line, C_p Statistic is minimized when the number of predictors is 3. Thus the best model selected with Mallows C_p is the same model selected with AIC: $$ gamble = 24.139 + 4.898 * income - 22.960 * sex - 2.747 * verbal $$*

```{r, fig.width=4, fig.height=3}
plot(rs$cp ~ I(1:p), xlab="Number of Predictors",
ylab=expression(paste(C[p], " Statistic")))
abline(1,1)
which.min(rs$cp)
```