extbf {Solution to BIOSTAT 285 Spring 2020 Homework 1}

Nan Liu

*Remark.* For **Computational Part**, please complete your answer in the **RMarkdown** file and summit the generated PDF and RMD files. Related packages have been loaded in setup.

## Computational Part

1. (Model Selection, [ISL] 6.8, *25 pt*) In this exercise, we will generate simulated data, and will then use this data to perform model selection.
   1. Use the rnorm function to generate a predictor of length , as well as a noise vector of length .

* **solution**:

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

set.seed(231)  
x <- rnorm(100,2,2)  
epsilon <- rnorm(100)

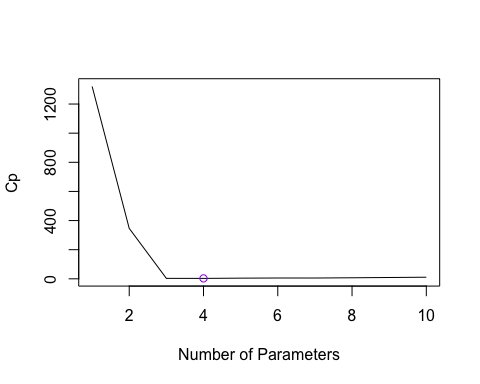
(b) Generate a response vector $\bm{Y}$ of length $n = 100$ according to the model $$ Y = \beta\_0 + \beta\_1 X + \beta\_2 X^2 + \beta\_3 X^3 + \epsilon, $$ where $\beta\_0 = 3$, $\beta\_1 = 2$, $\beta\_2 = -3$, $\beta\_3 = 0.3$.

y <- 3 + 2\*x -3\*x^2 + 0.3\*x^3 + epsilon

(c) Use the `regsubsets` function from `leaps` package to perform best subset selection in order to choose the best model from the set of predictors $(X, X^2, \cdots, X^{10})$. What are the best models obtained according to $C\_p$, BIC, and adjusted $R^2$, respectively? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained.

library(leaps)  
leaps <- regsubsets(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7)+I(x^8)+I(x^9)+I(x^10),data=data.frame(x=x,y=y),nvmax=10)  
summary <- summary(leaps)

#Cp  
plot(1:10, summary$cp, xlab="Number of Parameters", ylab="Cp", type = "l")  
min\_cp <- min(summary$cp)  
points(c(1:10)[summary$cp==min\_cp], min\_cp, pch=1, col="purple")

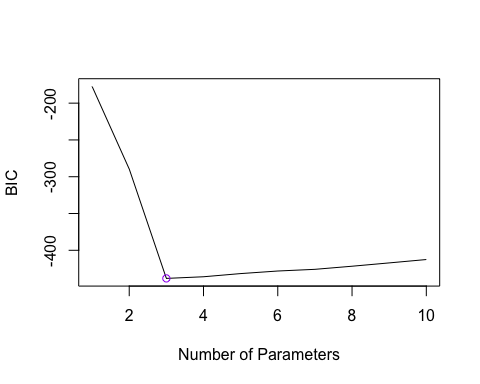


#report the coefficients  
lm1<-lm(y~I(x)+I(x^2)+I(x^3)+I(x^5), data=data.frame(x=x,y=y) )  
summary(lm1)

##   
## Call:  
## lm(formula = y ~ I(x) + I(x^2) + I(x^3) + I(x^5), data = data.frame(x = x,   
## y = y))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2618 -0.6570 -0.0308 0.6874 2.4294   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4312183 0.1867907 18.369 <2e-16 \*\*\*  
## I(x) 1.8246339 0.1348067 13.535 <2e-16 \*\*\*  
## I(x^2) -3.0608067 0.0617377 -49.578 <2e-16 \*\*\*  
## I(x^3) 0.3309170 0.0219861 15.051 <2e-16 \*\*\*  
## I(x^5) -0.0006367 0.0004219 -1.509 0.135   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.055 on 95 degrees of freedom  
## Multiple R-squared: 0.9898, Adjusted R-squared: 0.9894   
## F-statistic: 2315 on 4 and 95 DF, p-value: < 2.2e-16

The best models obtained according to is:

#BIC  
plot(1:10, summary$bic, xlab="Number of Parameters", ylab="BIC", type = "l")  
min\_bic <- min(summary$bic)  
points(c(1:10)[summary$bic==min\_bic], min\_bic, pch=1, col="purple")

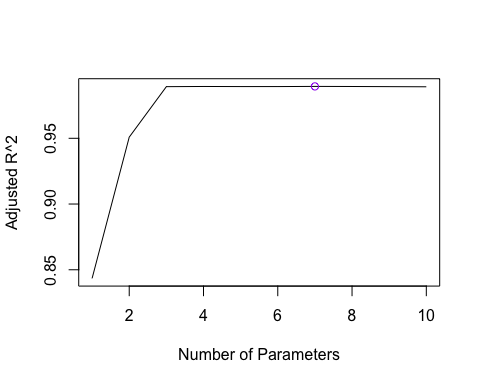


#report the coefficients  
lm2<-lm(y~I(x)+I(x^2)+I(x^3), data=data.frame(x=x,y=y) )  
summary(lm2)

##   
## Call:  
## lm(formula = y ~ I(x) + I(x^2) + I(x^3), data = data.frame(x = x,   
## y = y))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.38493 -0.69473 0.02238 0.76729 2.36562   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.348516 0.179755 18.63 <2e-16 \*\*\*  
## I(x) 1.954090 0.104679 18.67 <2e-16 \*\*\*  
## I(x^2) -3.019427 0.055681 -54.23 <2e-16 \*\*\*  
## I(x^3) 0.300989 0.009554 31.50 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.062 on 96 degrees of freedom  
## Multiple R-squared: 0.9896, Adjusted R-squared: 0.9893   
## F-statistic: 3046 on 3 and 96 DF, p-value: < 2.2e-16

The best models obtained according to BIC is:

#Adjusted R^2  
plot(1:10, summary$adjr2, xlab="Number of Parameters", ylab="Adjusted R^2", type = "l")  
max\_adjr2 <- max(summary$adjr2)  
points(c(1:10)[summary$adjr2==max\_adjr2], max\_adjr2, pch=1, col="purple")



#report the coefficients  
lm3<-lm(y~I(x)+I(x^2)+I(x^3)+I(x^4)+I(x^6)+I(x^7)+I(x^8), data=data.frame(x=x,y=y) )  
summary(lm3)

##   
## Call:  
## lm(formula = y ~ I(x) + I(x^2) + I(x^3) + I(x^4) + I(x^6) + I(x^7) +   
## I(x^8), data = data.frame(x = x, y = y))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.38214 -0.64580 -0.05089 0.72526 2.44031   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4373319 0.2067895 16.622 < 2e-16 \*\*\*  
## I(x) 1.4944353 0.3577604 4.177 6.71e-05 \*\*\*  
## I(x^2) -2.8801062 0.2246505 -12.820 < 2e-16 \*\*\*  
## I(x^3) 0.6509693 0.2009815 3.239 0.00167 \*\*   
## I(x^4) -0.2166670 0.1345038 -1.611 0.11064   
## I(x^6) 0.0238815 0.0139085 1.717 0.08934 .   
## I(x^7) -0.0053896 0.0030492 -1.768 0.08045 .   
## I(x^8) 0.0003548 0.0001971 1.800 0.07513 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.052 on 92 degrees of freedom  
## Multiple R-squared: 0.9902, Adjusted R-squared: 0.9895   
## F-statistic: 1330 on 7 and 92 DF, p-value: < 2.2e-16

The best models obtained according to adjusted is,

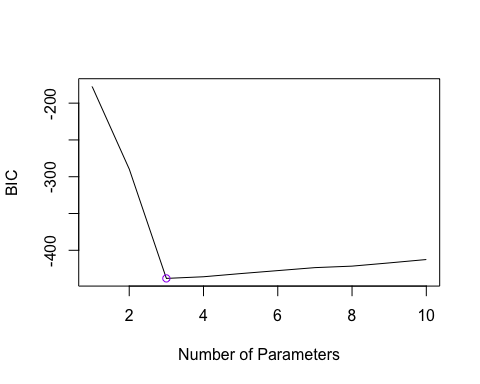
(d) Repeat (c), using forward stepwise selection and also using backward stepwise selection. How does your answer compare to the results in (c)?

#forward stepwise selection  
leaps1 <- regsubsets(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7)+I(x^8)+I(x^9)+I(x^10), data=data.frame(x=x,y=y),nvmax=10, method="forward")  
summary1 <- summary(leaps1)

```r  
#Cp  
plot(1:10, summary1$cp, xlab="Number of Parameters", ylab="Cp", type = "l")  
min\_cp1 <- min(summary1$cp)  
points(c(1:10)[summary1$cp==min\_cp1], min\_cp1, pch=1, col="purple")  
```  
  
![](hw1sol\_files/figure-docx/unnamed-chunk-8-1.png)<!-- -->  
  
```r  
#report the coefficients  
lm4<-lm(y~I(x)+I(x^2)+I(x^3)+I(x^5), data=data.frame(x=x,y=y) )  
summary(lm4)  
```  
  
```  
##   
## Call:  
## lm(formula = y ~ I(x) + I(x^2) + I(x^3) + I(x^5), data = data.frame(x = x,   
## y = y))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2618 -0.6570 -0.0308 0.6874 2.4294   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4312183 0.1867907 18.369 <2e-16 \*\*\*  
## I(x) 1.8246339 0.1348067 13.535 <2e-16 \*\*\*  
## I(x^2) -3.0608067 0.0617377 -49.578 <2e-16 \*\*\*  
## I(x^3) 0.3309170 0.0219861 15.051 <2e-16 \*\*\*  
## I(x^5) -0.0006367 0.0004219 -1.509 0.135   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.055 on 95 degrees of freedom  
## Multiple R-squared: 0.9898, Adjusted R-squared: 0.9894   
## F-statistic: 2315 on 4 and 95 DF, p-value: < 2.2e-16  
```

The best models obtained according to is:(same!!!)

#BIC  
plot(1:10, summary1$bic, xlab="Number of Parameters", ylab="BIC", type = "l")  
min\_bic1 <- min(summary1$bic)  
points(c(1:10)[summary1$bic==min\_bic1], min\_bic1, pch=1, col="purple")

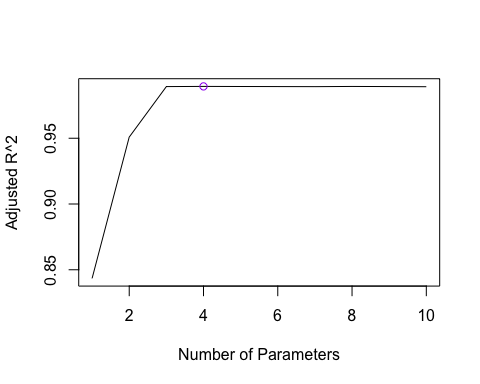


#report the coefficients  
lm5<-lm(y~I(x)+I(x^2)+I(x^3), data=data.frame(x=x,y=y) )  
summary(lm5)

##   
## Call:  
## lm(formula = y ~ I(x) + I(x^2) + I(x^3), data = data.frame(x = x,   
## y = y))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.38493 -0.69473 0.02238 0.76729 2.36562   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.348516 0.179755 18.63 <2e-16 \*\*\*  
## I(x) 1.954090 0.104679 18.67 <2e-16 \*\*\*  
## I(x^2) -3.019427 0.055681 -54.23 <2e-16 \*\*\*  
## I(x^3) 0.300989 0.009554 31.50 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.062 on 96 degrees of freedom  
## Multiple R-squared: 0.9896, Adjusted R-squared: 0.9893   
## F-statistic: 3046 on 3 and 96 DF, p-value: < 2.2e-16

The best models obtained according to BIC is:

#Adjusted R^2  
plot(1:10, summary1$adjr2, xlab="Number of Parameters", ylab="Adjusted R^2", type = "l")  
max\_adjr2\_1 <- max(summary1$adjr2)  
points(c(1:10)[summary1$adjr2==max\_adjr2\_1], max\_adjr2\_1, pch=1, col="purple")



#report the coefficients  
lm6<-lm(y~I(x)+I(x^2)+I(x^3)+I(x^4)+I(x^6)+I(x^7)+I(x^8), data=data.frame(x=x,y=y) )  
summary(lm6)

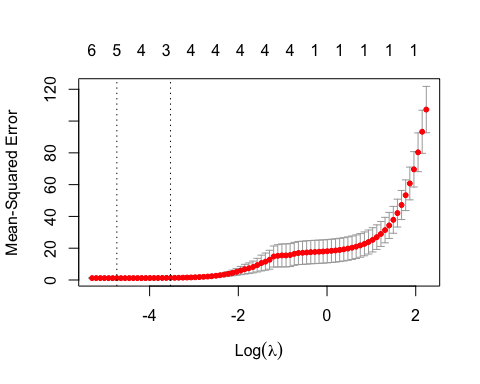
##   
## Call:  
## lm(formula = y ~ I(x) + I(x^2) + I(x^3) + I(x^4) + I(x^6) + I(x^7) +   
## I(x^8), data = data.frame(x = x, y = y))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.38214 -0.64580 -0.05089 0.72526 2.44031   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4373319 0.2067895 16.622 < 2e-16 \*\*\*  
## I(x) 1.4944353 0.3577604 4.177 6.71e-05 \*\*\*  
## I(x^2) -2.8801062 0.2246505 -12.820 < 2e-16 \*\*\*  
## I(x^3) 0.6509693 0.2009815 3.239 0.00167 \*\*   
## I(x^4) -0.2166670 0.1345038 -1.611 0.11064   
## I(x^6) 0.0238815 0.0139085 1.717 0.08934 .   
## I(x^7) -0.0053896 0.0030492 -1.768 0.08045 .   
## I(x^8) 0.0003548 0.0001971 1.800 0.07513 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.052 on 92 degrees of freedom  
## Multiple R-squared: 0.9902, Adjusted R-squared: 0.9895   
## F-statistic: 1330 on 7 and 92 DF, p-value: < 2.2e-16

The best models obtained according to adjusted is,

The answers are the same as those results in (c).

(e) Now fit a LASSO model with `glmnet` function from `glmnet` package to the simulated data, again using $(X,X^2,\cdots,X^{10})$ as predictors. Use cross-validation to select the optimal value of $\lambda$. Create plots of the cross-validation error as a function of $\lambda$. Report the resulting coefficient estimates, and discuss the results obtained.

library(glmnet)  
x1=cbind(x,x^2,x^3,x^4,x^5,x^6,x^7,x^8,x^9,x^10)  
y=y  
# Cross-validation to select lambda  
lasso.cv = cv.glmnet(x1,y, alpha=1)  
lad<- lasso.cv$lambda.min  
plot(lasso.cv)



# coefficient estimate when refitting model with lamda.min  
lamodel=glmnet(x1,y,alpha=1, lambda=lad)  
summary(lamodel)

## Length Class Mode   
## a0 1 -none- numeric  
## beta 10 dgCMatrix S4   
## df 1 -none- numeric  
## dim 2 -none- numeric  
## lambda 1 -none- numeric  
## dev.ratio 1 -none- numeric  
## nulldev 1 -none- numeric  
## npasses 1 -none- numeric  
## jerr 1 -none- numeric  
## offset 1 -none- logical  
## call 5 -none- call   
## nobs 1 -none- numeric

coef(lamodel)

## 11 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 3.341576e+00  
## x 1.885177e+00  
## -3.005461e+00  
## 3.040626e-01  
## .   
## .   
## .   
## .   
## -2.191847e-07  
## -6.097948e-08  
## -5.256913e-09

修改！The plot includes the cross-validation curve (red dotted line), and upper and lower stan- dard deviation curves along the sequence of λ values. Two selected λ values are indicated by the vertical dotted line: lambda giving the minimum cv error and the lambda within one standard devation of the minimum cv error: in this example, they are 98.97694 and 108.6271 respectively. With the value of λ giving the minimum cv error, the Lasso shrinks the majority predic- tors to zero, and only leaves X2 and X3 nozero.

(f) Now generate a response vector $Y$ according to the model $$Y = \beta\_0 + \beta\_7 X^7 + \epsilon,$$ where $\beta\_7 = 7$, and perform best subset selection and the LASSO. Discuss the results obtained.

1. (Prediction, [ISL] 6.9, *20 pt*) In this exercise, we will predict the number of applications received (Apps) using the other variables in the College data set from ISLR package.
   1. Randomly split the data set into equal sized training set and test set (1:1).
   2. Fit a linear model using least squares on the training set, and report the test error obtained.
   3. Fit a ridge regression model on the training set, with chosen by 5-fold cross-validation. Report the test error obtained.
   4. Fit a LASSO model on the training set, with chosen by 5-fold cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.
   5. Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these three approaches?