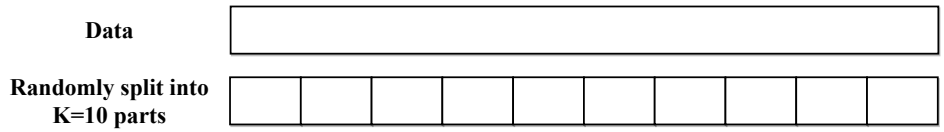
ESI 4606 Analytics I - Foundations of Data Science
Homework 5
Due: Nov. 9st (11:00AM), 2022

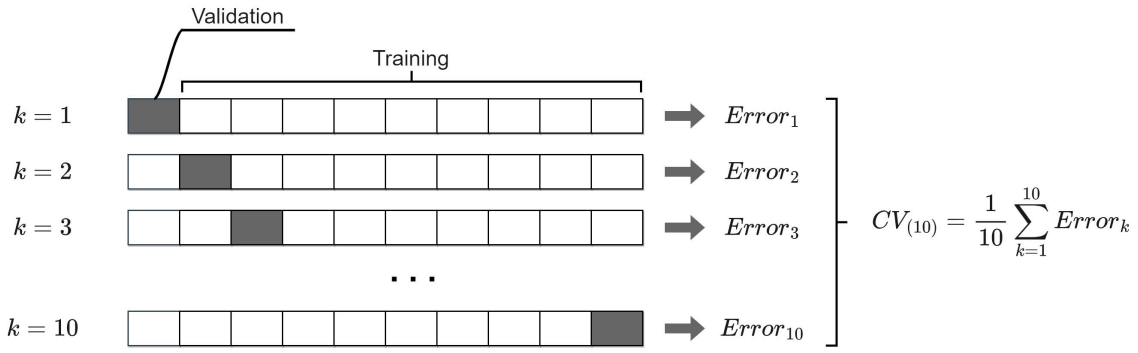
Problem 1 (1 point)

Explain how to utilize 10-fold cross-validation to estimate test error of multiple linear regression. (Use words, figures and mathematical notations to provide a clear description)

Solution: (a) Step 1: randomly split data into 10 roughly equal parts:



Step 2: for k=1,...,10, repeat the following: Leave the k^{th} portion out, and train the multiple linear regression using the other 9 parts. Calculate the cross-validation error (i.e., mean square error) on the k^{th} portion as $Error_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(Y_i - \hat{Y}_i)^2$, where n_k is the number of observations in the k^{th} portion, $y_{i,k}$ and $\hat{y}_{i,k}, i = 1,...,n_k$ are actual classes and predicted classes for observations in the k^{th} portion.



Step 3: Compute cross-validation error as $CV_{(10)} = \frac{1}{10}\sum_{k=1}^{10} Error_k$.

Problem 2 (3 points)

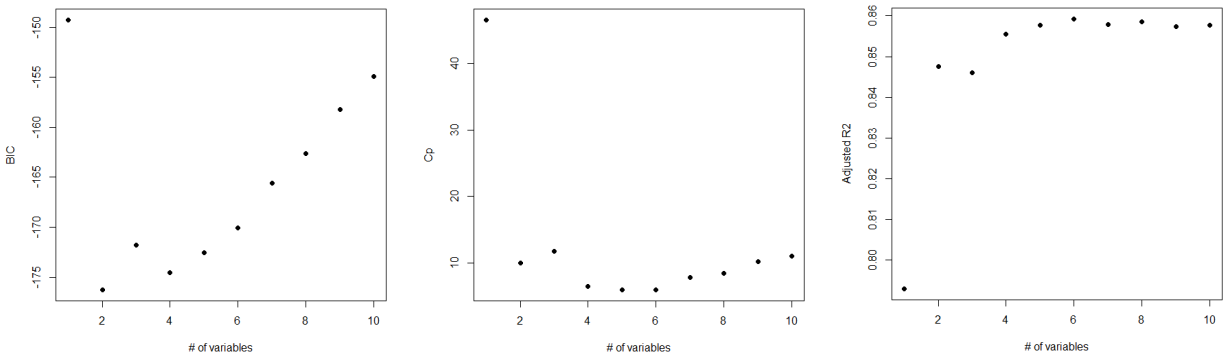
Solution: (a) Figure 1 shows the best subset selection results.

When using BIC for model selection, Figure 2a shows that the best model (with

the minimum BIC value) includes 2 predictors, namely x_1, x_2 shown in Figure 1. Figure 3a reports the corresponding estimated coefficients.
When using C_p for model selection, Figure 2b shows that the best model (with the minimum C_p value) includes 5 predictors, namely $x_2, x_3, x_7, x_8, x_{10}$ shown in Figure 1. Figure 3b reports the corresponding estimated coefficients.
When using adjusted R^2 for model selection, Figure 2c shows that the best model (with the adjusted R^2 value) includes 6 predictors, namely $x_1, x_2, x_3, x_5, x_7, x_9$ shown in Figure 1. Figure 3c reports the corresponding estimated coefficients.

```
Selection Algorithm: exhaustive
          x1  x2  x3  x4  x5  x6  x7  x8  x9  x10
1  ( 1 )  " " "*" " " " " " " " " " " " " " " " "
2  ( 1 )  "*" "*" " " " " " " " " " " " " " " " "
3  ( 1 )  "*" "*" " " " " " " " " "*" " " " " " "
4  ( 1 )  " " "*" " " " " "*" " " "*" " " "*" " "
5  ( 1 )  " " "*" "*" " " " " " " "*" "*" " " "*"
6  ( 1 )  "*" "*" "*" " " " " "*" " " "*" " " "*" " "
7  ( 1 )  "*" "*" "*" " " " " "*" " " "*" "*" " " "*"
8  ( 1 )  "*" "*" "*" " " " " "*" "*" "*" "*" "*" " "
9  ( 1 )  "*" "*" "*" " " " " "*" "*" "*" "*" "*" "*"
10  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

Figure 1: Variable selection results of best subset selection



(a) BIC (b) C_p (c) Adjusted R^2

Figure 2: Variable selection criteria using best subset selection

(b) Figure 4 shows the forward stepwise selection results.
When using BIC for model selection, Figure 5a shows that the best model (with the minimum BIC value) includes 2 predictors, namely x_1, x_2 shown in Figure 4. Figure 6a reports the corresponding estimated coefficients.
When using C_p for model selection, Figure 5b shows that the best model (with the minimum C_p value) includes 6 predictors, namely $x_1, x_2, x_5, x_6, x_7, x_{10}$ shown in Figure 4. Figure 6b reports the corresponding estimated coefficients.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96281    0.09251  10.407  < 2e-16 ***
x1           0.56270    0.09369   6.006 3.31e-08 ***
x2          -2.18096    0.09369 -23.280  < 2e-16 ***
```

(a) Selected variables based on BIC

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96281    0.08935  10.776  < 2e-16 ***
x2          -1.90297    0.21385  -8.899 4.03e-14 ***
x3           1.50748    0.25911   5.818 8.19e-08 ***
x7          -1.68346    0.38939  -4.323 3.82e-05 ***
x8          -3.89877    1.44648  -2.695  0.00833 **
x10          4.37547    1.42908   3.062  0.00287 **
```

(b) Selected variables based on C_p

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96281    0.08887  10.834  < 2e-16 ***
x1           1.04303    0.48848   2.135  0.03537 *
x2          -2.18226    0.10585 -20.617  < 2e-16 ***
x3          -6.58106    3.40468  -1.933  0.05629 .
x5          25.77460   10.03529   2.568  0.01181 *
x7         -36.81267   12.53432  -2.937  0.00418 **
x9          17.22748    5.51293   3.125  0.00237 **
```

(c) Selected variables based on R^2

Figure 3: Estimated coefficients summary based on different variable selection criteria using best subset selection

When using adjusted R^2 for model selection, Figure 5c shows that the best model (with the maximum adjusted R^2 value) includes 6 predictors, namely x_1, x_2, x_5, x_6, x_7, x_{10} shown in Figure 4. Figure 6c reports the corresponding estimated coefficients.

```
Selection Algorithm: forward
          x1  x2  x3  x4  x5  x6  x7  x8  x9  x10
1  ( 1 )  " " "*" " " " " " " " " " " " " " " " "
2  ( 1 )  "*" "*" " " " " " " " " " " " " " " " "
3  ( 1 )  "*" "*" " " " " " " " " " " "*" " " " "
4  ( 1 )  "*" "*" " " " " " " " " " " "*" " " "*"
5  ( 1 )  "*" "*" " " " " " " " " "*" "*" " " "*"
6  ( 1 )  "*" "*" " " " " " "*" "*" "*" " " " "*"
7  ( 1 )  "*" "*" " " "*" "*" "*" "*" " " " "*"
8  ( 1 )  "*" "*" "*" "*" "*" "*" "*" "*" " " " "
9  ( 1 )  "*" "*" "*" "*" "*" "*" "*" "*" "*" " "
10  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

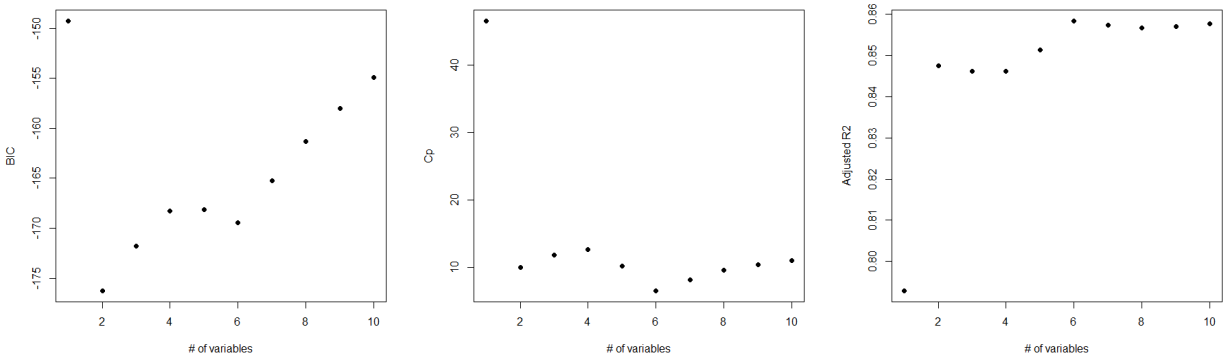
Figure 4: Variable selection results of forward stepwise selection

(c) Figure 7 shows the backward stepwise selection results.

When using BIC for model selection, Figure 8a shows that the best model (with the minimum BIC value) includes 4 predictors, namely x_2, x_5, x_7, x_9 shown in Figure 7. Figure 9a reports the corresponding estimated coefficients.

When using C_p for model selection, Figure 8b shows that the best model (with the minimum C_p value) includes 6 predictors, namely $x_1, x_2, x_3, x_5, x_7, x_9$ shown in Figure 7. Figure 9b reports the corresponding estimated coefficients.

When using adjusted R^2 for model selection, Figure 8c shows that the best model



(a) BIC (b) C_p (c) Adjusted R^2

Figure 5: Variable selection criteria using forward stepwise selection

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96281    0.09251  10.407  < 2e-16 ***
x1           0.56270    0.09369   6.006 3.31e-08 ***
x2          -2.18096    0.09369 -23.280  < 2e-16 ***
```

(a) Selected variables based on BIC

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96281    0.08914  10.801  < 2e-16 ***
x1           0.35756    0.17906   1.997  0.04876 *
x2          -1.70398    0.26440  -6.445 5.08e-09 ***
x5           2.49725    1.04584   2.388  0.01897 *
x6          -2.22214    0.79423  -2.798  0.00625 **
x7          -3.07218    1.13934  -2.696  0.00832 **
x10          2.57529    0.78834   3.267  0.00152 **
```

(b) Selected variables based on C_p

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96281    0.08914  10.801  < 2e-16 ***
x1           0.35756    0.17906   1.997  0.04876 *
x2          -1.70398    0.26440  -6.445 5.08e-09 ***
x5           2.49725    1.04584   2.388  0.01897 *
x6          -2.22214    0.79423  -2.798  0.00625 **
x7          -3.07218    1.13934  -2.696  0.00832 **
x10          2.57529    0.78834   3.267  0.00152 **
```

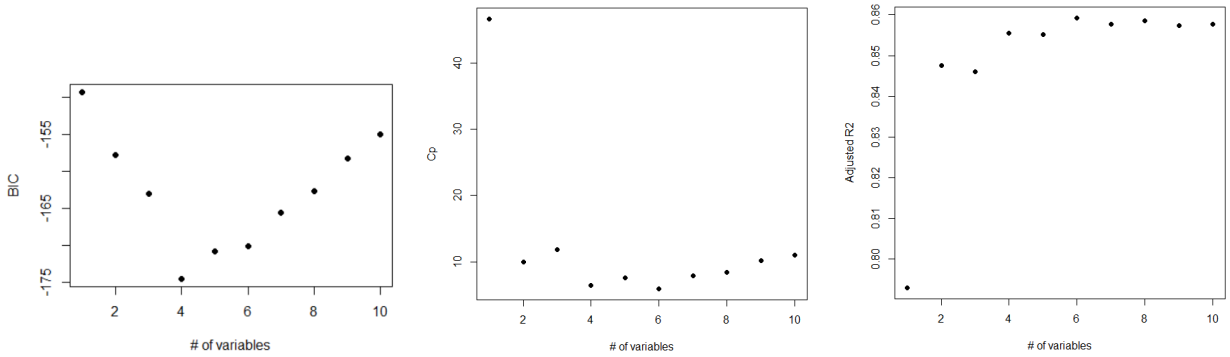
(c) Selected variables based on R^2

Figure 6: Estimated coefficients summary based on different variable selection criteria using forward stepwise selection

(with the maximum adjusted R^2 value) includes 6 predictors, namely x_1, x_2, x_3, x_5, x_7, x_9 shown in Figure 7. Figure 9c reports the corresponding estimated coefficients.

```
Selection Algorithm: backward
          x1  x2  x3  x4  x5  x6  x7  x8  x9  x10
1  ( 1 )  " " "*" " " " " " " " " " " " " " " " "
2  ( 1 )  " " "*" " " " " " " "*" " " " " " " " "
3  ( 1 )  " " "*" " " " " " " "*" " " "*" " " " "
4  ( 1 )  " " "*" " " " " " " "*" " " "*" " " "*"
5  ( 1 )  "*" "*" " " " " " " "*" " " "*" " " "*"
6  ( 1 )  "*" "*" "*" " " " " "*" " " "*" " " "*"
7  ( 1 )  "*" "*" "*" " " " " "*" " " "*" "*" "*"
8  ( 1 )  "*" "*" "*" " " " " "*" "*" "*" "*" "*"
9  ( 1 )  "*" "*" "*" " " " " "*" "*" "*" "*" "*"
10  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

Figure 7: Variable selection results of backward stepwise selection



(a) BIC (b) C_p (c) Adjusted R^2

Figure 8: Variable selection criteria using forward stepwise selection

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.96281    0.09006  10.691  < 2e-16 ***
x2           -2.19693    0.10685 -20.562  < 2e-16 ***
x5            8.60846    1.67875   5.128 1.54e-06 ***
x7          -17.69143    3.94104  -4.489 2.01e-05 ***
x9            9.65186    2.36675   4.078 9.44e-05 ***
```

(a) Selected variables based on BIC

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.96281    0.08887  10.834  < 2e-16 ***
x1            1.04303    0.48848   2.135  0.03537 *
x2           -2.18226    0.10585 -20.617  < 2e-16 ***
x3           -6.58106    3.40468  -1.933  0.05629 .
x5           25.77460   10.03529   2.568  0.01181 *
x7          -36.81267   12.53432  -2.937  0.00418 **
x9           17.22748    5.51293   3.125  0.00237 **
```

(b) Selected variables based on C_p

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.96281    0.08887  10.834  < 2e-16 ***
x1            1.04303    0.48848   2.135  0.03537 *
x2           -2.18226    0.10585 -20.617  < 2e-16 ***
x3           -6.58106    3.40468  -1.933  0.05629 .
x5           25.77460   10.03529   2.568  0.01181 *
x7          -36.81267   12.53432  -2.937  0.00418 **
x9           17.22748    5.51293   3.125  0.00237 **
```

(c) Selected variables based on R^2

Figure 9: Estimated coefficients summary based on different variable selection criteria using backward stepwise selection

Problem 3 (1 point)

Let $p = 9$ and $n = 50$, where p is the number of predictors and n is the sample size. When linear regression is considered for data fitting, answer the following

questions.

(a) To determine the best model by selecting the best subset of relevant predictors, how many models in total need to be compared if best subset selection method is used?

Solution: To determine the best subset of predictors, there will be 2 possible decisions for each of the 9 predictors: to select it or not to select it. So the total number of possibilities is $2^9 = 512$. These will include a model with no predictor. You can consider it as a linear model with only β_0, and in this case the answer is 512. Or you can address that a constant is not a linear model, so your answer is 511. Both explanations are reasonable.

(b) To determine M_3, i.e., best model with 3 predictors, how many models with 3 predictors need to be compared if best subset selection method is used?

Solution: It's to choose 3 out of 9 predictors. The number of possibilities $= \binom{9}{3} = 84$.

(c) To determine M_3, i.e., best model with 3 predictors, how many models with 3 predictors need to be compared if forward stepwise selection method is used?

Solution: This is to say that given $M2$, i.e., best model with 2 predictors, use forward stepwise to select the third predictor. So the number of possibilities $= 9-2 = 7$.

(d) To determine M_3, i.e., best model with 3 predictors, how many models with 3 predictors need to be compared if backward stepwise selection method is used?

Solution: This is to exclude a predictor from $M4$ (best model with 4 predictors). So the number of possible selections is 4.

Appendix A: R codes for Problem 2

```
#import "HM5.txt" and call it "data"
library(leaps)
#best subset selection
lm.exhaustive=regsubsets(y~.,data=data,nvmax=10)
summary(lm.exhaustive)
plot(summary(lm.exhaustive)$bic,xlab="# of variables", ylab="BIC",pch=19)
which.min(summary(lm.exhaustive)$bic)
summary(lm(y~x1+x2,data=data))
plot(summary(lm.exhaustive)$cp,xlab="# of variables", ylab="Cp",pch=19)
which.min(summary(lm.exhaustive)$cp)
summary(lm(y~x2+x3+x7+x8+x10,data=data))
plot(summary(lm.exhaustive)$adjr2,xlab="# of variables", ylab="Adjusted R2",pch=19)
which.max(summary(lm.exhaustive)$adjr2)
summary(lm(y~x1+x2+x3+x5+x7+x9,data=data))
#forward stepwise selection
lm.forward=regsubsets(y~.,data=data,nvmax=10,method="forward")
summary(lm.forward)
plot(summary(lm.forward)$bic,xlab="# of variables", ylab="BIC",pch=19)
which.min(summary(lm.forward)$bic)
summary(lm(y~x1+x2,data=data))
plot(summary(lm.forward)$cp,xlab="# of variables", ylab="Cp",pch=19)
which.min(summary(lm.forward)$cp)
summary(lm(y~x1+x2+x5+x6+x7+x10,data=data))
plot(summary(lm.forward)$adjr2,xlab="# of variables", ylab="Adjusted R2",pch=19)
which.max(summary(lm.forward)$adjr2)
summary(lm(y~x1+x2+x5+x6+x7+x10,data=data))
#bacward stepwise selection
lm.backward=regsubsets(y~.,data=data,nvmax=10,method="backward")
summary(lm.backward)
plot(summary(lm.backward)$bic,xlab="# of variables", ylab="BIC",pch=19)
which.min(summary(lm.backward)$bic)
summary(lm(y~x2+x5+x7+x9,data=data))
plot(summary(lm.backward)$cp,xlab="# of variables", ylab="Cp",pch=19)
which.min(summary(lm.backward)$cp)
summary(lm(y~x1+x2+x3+x5+x7+x9,data=data))
plot(summary(lm.backward)$adjr2,xlab="# of variables", ylab="Adjusted R2",pch=19)
which.max(summary(lm.backward)$adjr2)
summary(lm(y~x1+x2+x3+x5+x7+x9,data=data))
```