

Stat154_Quiz3

Samba Njie Jr

October 16, 2016

Housing Values in Suburbs of Boston

Dataset Description: `housing.csv` concerns housing values in suburbs of Boston. The dataset was created by Harrison, D. and Rubinfeld, D.L. and analyzed in ‘*Hedonic prices and the demand for clean air*’, *J. Environment Economics and Management*, vol. 5, 81 - 102, 1978. There are 506 observations and 12 continuous attributes including the response variable MEDV.

Attribute Information:

- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq. ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per \$10,000
- PTRATIO: pupil-teacher ratio by town
- LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes in \$1000's.

Tasks:

1. Create randomly sampled training and test sets from the dataset using 90% of the observations for training and 10% for testing. Put aside your test set and only use it for the last task.

First, we load the libraries:

```
library(ISLR)
library(MASS)
library(dplyr)
library(utils)
library(corrplot)
library(glmnet)
```

Read in the housing.csv data set:

```
dataset = read.csv(file = "/Users/sambamamba/Documents/154_files/housing.csv")
head(dataset) #inspect the columns and rows in the beginning
```

```
##   X    CRIM ZN  INDUS CHAS    NOX     RM   AGE     DIS  RAD  TAX  PTRATIO  LSTAT
## 1 1 0.00632 18   2.31    0 0.538 6.575 65.2 4.0900    1  296    15.3   4.98
## 2 2 0.02731  0   7.07    0 0.469 6.421 78.9 4.9671    2  242    17.8   9.14
## 3 3 0.02729  0   7.07    0 0.469 7.185 61.1 4.9671    2  242    17.8   4.03
## 4 4 0.03237  0   2.18    0 0.458 6.998 45.8 6.0622    3  222    18.7   2.94
## 5 5 0.06905  0   2.18    0 0.458 7.147 54.2 6.0622    3  222    18.7   5.33
## 6 6 0.02985  0   2.18    0 0.458 6.430 58.7 6.0622    3  222    18.7   5.21
##   MEDV
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

Now let us set up randomly sampled training and test sets, 90% training, 10% testing.

```
set.seed(1)
n_obs = nrow(dataset) #number of observations
n_features = ncol(dataset) #number of features

prop_training = round(0.90*n_obs) #subset of observations composing the training set
prop_test = round(0.10*n_obs) #subset of observations for test set

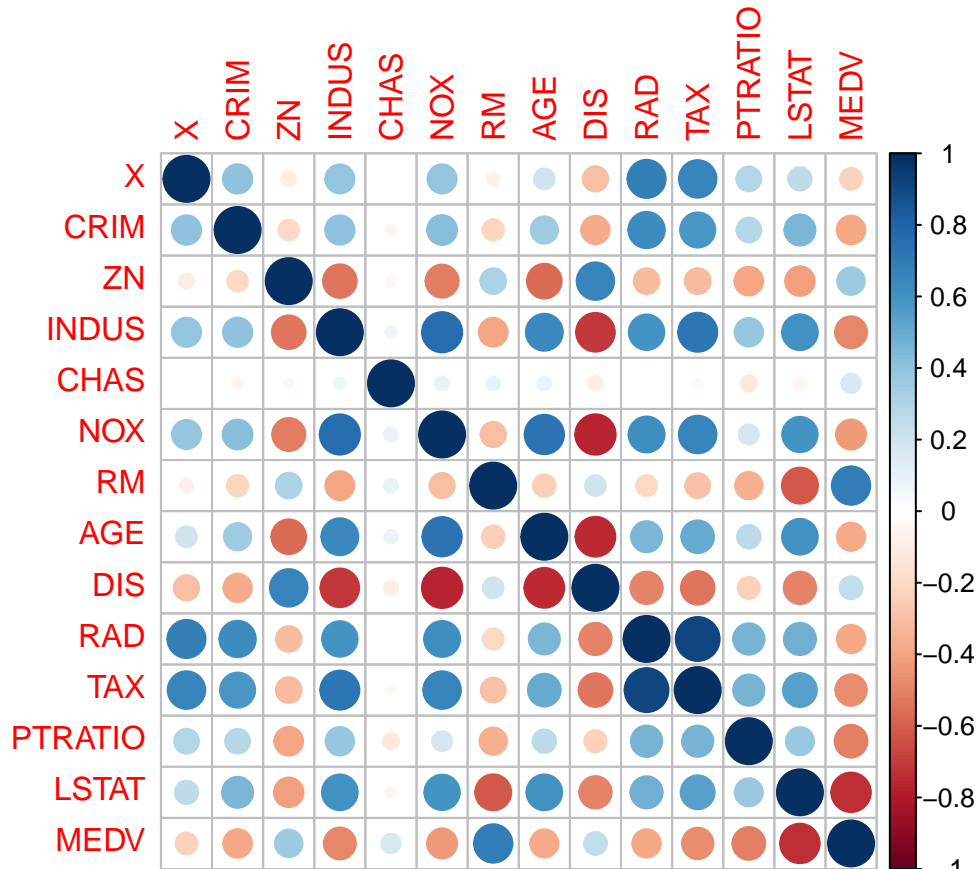
training_index = sample(n_obs, prop_training) #samples integers from 1 to n_obs

test_index = dataset$X[-training_index]

train_set <- dataset %>%
  subset(X %in% training_index)

test_set <- dataset %>%
  subset(!(X %in% training_index))
```

2. Plot the correlation matrix of all attributes. Which attributes you deem more predictive of the housing prices?



##	X	CRIM	ZN	INDUS	CHAS
## X	1.000000000	0.40740717	-0.10339336	0.39943885	-0.003759115
## CRIM	0.407407172	1.00000000	-0.20046922	0.40658341	-0.055891582
## ZN	-0.103393357	-0.20046922	1.00000000	-0.53382819	-0.042696719
## INDUS	0.399438850	0.40658341	-0.53382819	1.00000000	0.062938027
## CHAS	-0.003759115	-0.05589158	-0.04269672	0.06293803	1.00000000
## NOX	0.398736174	0.42097171	-0.51660371	0.76365145	0.091202807
## RM	-0.079971150	-0.21924670	0.31199059	-0.39167585	0.091251225
## AGE	0.203783510	0.35273425	-0.56953734	0.64477851	0.086517774
## DIS	-0.302210959	-0.37967009	0.66440822	-0.70802699	-0.099175780
## RAD	0.686001976	0.62550515	-0.31194783	0.59512927	-0.007368241
## TAX	0.666625924	0.58276431	-0.31456332	0.72076018	-0.035586518
## PTRATIO	0.291074227	0.28994558	-0.39167855	0.38324756	-0.121515174
## LSTAT	0.258464770	0.45562148	-0.41299457	0.60379972	-0.053929298
## MEDV	-0.226603643	-0.38830461	0.36044534	-0.48372516	0.175260177
##	NOX	RM	AGE	DIS	RAD
## X	0.39873617	-0.07997115	0.20378351	-0.30221096	0.686001976
## CRIM	0.42097171	-0.21924670	0.35273425	-0.37967009	0.625505145
## ZN	-0.51660371	0.31199059	-0.56953734	0.66440822	-0.311947826
## INDUS	0.76365145	-0.39167585	0.64477851	-0.70802699	0.595129275
## CHAS	0.09120281	0.09125123	0.08651777	-0.09917578	-0.007368241
## NOX	1.00000000	-0.30218819	0.73147010	-0.76923011	0.611440563
## RM	-0.30218819	1.00000000	-0.24026493	0.20524621	-0.209846668
## AGE	0.73147010	-0.24026493	1.00000000	-0.74788054	0.456022452
## DIS	-0.76923011	0.20524621	-0.74788054	1.00000000	-0.494587930
## RAD	0.61144056	-0.20984667	0.45602245	-0.49458793	1.00000000

```
## TAX      0.66802320 -0.29204783  0.50645559 -0.53443158  0.910228189
## PTRATIO  0.18893268 -0.35550149  0.26151501 -0.23247054  0.464741179
## LSTAT    0.59087892 -0.61380827  0.60233853 -0.49699583  0.488676335
## MEDV     -0.42732077  0.69535995 -0.37695457  0.24992873 -0.381626231
##          TAX      PTRATIO      LSTAT      MEDV
## X          0.66662592  0.2910742  0.2584648 -0.2266036
## CRIM       0.58276431  0.2899456  0.4556215 -0.3883046
## ZN        -0.31456332 -0.3916785 -0.4129946  0.3604453
## INDUS     0.72076018  0.3832476  0.6037997 -0.4837252
## CHAS      -0.03558652 -0.1215152 -0.0539293  0.1752602
## NOX       0.66802320  0.1889327  0.5908789 -0.4273208
## RM        -0.29204783 -0.3555015 -0.6138083  0.6953599
## AGE       0.50645559  0.2615150  0.6023385 -0.3769546
## DIS       -0.53443158 -0.2324705 -0.4969958  0.2499287
## RAD       0.91022819  0.4647412  0.4886763 -0.3816262
## TAX       1.00000000  0.4608530  0.5439934 -0.4685359
## PTRATIO   0.46085304  1.0000000  0.3740443 -0.5077867
## LSTAT     0.54399341  0.3740443  1.0000000 -0.7376627
## MEDV     -0.46853593 -0.5077867 -0.7376627  1.0000000
```

Based on the correlation matrix visualized by calling `corr_matrix`, we can notice that matrix entries with darker shades of blue and red are heavily correlated with one another. The correlation coefficients are displayed in a matrix under the `corr_variables` object, and `corr_matrix` graphically displays these correlations, with the size and color glyphs representing the strength of the correlation. With the housing prices variable, `MEDV` being predicted, we can then infer that those glyphs or attributes that have the largest circles describe features that have the strongest correlation coefficients with the housing prices variable `MEDV`, with a strong blue color indicating positive correlation and a strong negative color indicating negative correlation.

From the data visualization we generated from the `corrplot()` function, we can presume that `LSTAT` and `INDUS` variables have strong negative correlations with `MEDV`, while `RM` and `ZN` have strong positive correlations. With these presumptions, we can witness if they hold true:

```
head(corr_matrix[14,])
```

```
##          X      CRIM      ZN      INDUS      CHAS      NOX
## -0.2266036 -0.3883046  0.3604453 -0.4837252  0.1752602 -0.4273208
```

Quantitatively, `LSTAT` and `PTRATIO` have the highest negative correlations, -0.737 and -0.508 respectively, and `RM` and `TAX` seem to have some of the highest positive correlations.

So for our hypotheses, we suppose that `RM` and `LSTAT` are most predictive of housing prices.

3. Implement Algorithm 6.1 and report the best model under C_p , BIC, adjusted R^2 and Cross-Validation (k-fold, k of your choice).

Algorithm 6.1 in the textbook is *best subset selection*, a subset selection method which will allow us to pick the best model to describe the data. Best subset selection fits a model for each possible combination of p predictors. The algorithm is as follows:

- (a) Let M_0 denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
- (b) For $k = 1, 2, \dots, p$:

- i. Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - ii. Pick the best among these $\binom{p}{k}$ models, and call it M_k . Here the *best* is defined as having the smallest RSS, or equivalently largest R^2 .
- (c) Select a single best model from among M_0, \dots, M_p using cross-validation, C_p , AIC, BIC, or adjusted R^2 .

Now for the implementation:

(a)

```
sum(is.na(dataset)) #determines number of NA values. Since 0, nothing to omit.
```

```
## [1] 0
```

```
response <- train_set$MEDV
```

```
M_0 <- mean(response) #null model, 0 predictors, and the RSS is simply the mean of the response variable
M_0
```

```
## [1] 22.4444
```

(b)

```
subsets <- function(p) { #finds the number of possible subsets of any sized set
  #p = ncol(data)
  total <- list()

  for (i in 1:p) {
    sets <- combn(p, i)
    total[[i]] <- sets
  }
  return(total)
}
```

Now we use the helper function `subsets` to index all the possible types of combinations of predictors, which is mathematically defined as the power set of a set of values, and fit each combination to a model. Hence we define `power_set`, a function that inputs a data frame and outputs a list of list of data frames labeled `sets`: each sublist is the M_k set of models.

```
power_set <- function(data, response_var) {
  #create empty lists sets and each_subset, which provides the RSS of every model of every list of list
  sets <- list()
  each_subset <- list()
  new_vec <- data.frame(rep(NA, nrow(data)))
  p <- ncol(data)

  powers <- subsets(p)
```

```

for (i in 1:length(powers)) {
  for (j in 1:ncol(powers[[i]]) ) {
    for (k in 1:length(powers[[i]][,j]) ) {
      t = powers[[i]][k, j]
      vec <- data.frame(data[t])
      new_vec <- cbind(new_vec, vec)
    }
    new_vec <- new_vec[-1] #eliminates NA list
    new_vec <- cbind(response_var, new_vec)
    models <- summary(lm(response_var ~ ., data = new_vec))
    each_subset[[j]] <- models #creates a list of j lists for each k model
  }
  sets[[i]] <- each_subset
}
sets <- na.omit(sets)
return(sets)
}

power_rss <- function(data, response_var) {
  #create empty lists sets and each_subset, which provides the RSS of every model of every list of list.
  sets <- list()
  each_subset <- list()
  new_vec <- data.frame(rep(NA, nrow(data)))
  p <- ncol(data)

  powers <- subsets(p)

  for (i in 1:length(powers)) {
    for (j in 1:ncol(powers[[i]]) ) {
      for (k in 1:length(powers[[i]][,j]) ) {
        t = powers[[i]][k, j]
        vec <- data.frame(data[t])
        new_vec <- cbind(new_vec, vec)
      }
      new_vec <- new_vec[-1] #eliminates NA list
      new_vec <- cbind(response_var, new_vec)
      models <- summary(lm(response_var ~ ., data = new_vec))[[6]]
      each_subset[[j]] <- models #creates a list of j lists for each k model
    }
    sets[[i]] <- each_subset
  }
  sets <- na.omit(sets)
  return(sets)
}

best_rss <- function(predictor_space, response) {
  bss <- power_rss(predictor_space, response)
  min_rss <- list()

  for (i in 1:length(bss)) {
    bss_vec <- rep(NA, length(bss[i]))
  }
}

```

```

    for (j in 1:length(bss[i])) { #turns every element of the power set to be a vector
      bss_vec[j] <- as.numeric(bss[[i]][j])
    }
    min_rss <- min(bss_vec)
  }
  return(min_rss)
}

index <- function(vec, value) {#indexes a set
  for (i in 1:length(vec)) {
    if (value == vec[i]) {
      yes <- i
    }
  }
  return(yes)
}

idx_best <- function(predictor_space, response) {#list of indices of each M_k with each value being the
  idx <- list()
  p <- ncol(predictor_space)
  brss <- best_rss(predictor_space, response)
  bss <- power_rss(predictor_space, response)

  for (i in 1:length(bss)) {
    bss_vec <- rep(NA, length(bss[i]))
    for (j in 1:length(bss[i])) { #turns every element of the power set to be a vector
      bss_vec[j] <- bss[[i]][j]
    }
    M_vecs[[i]] <- bss_vec
    idx[[i]] <- index(M_vecs[[i]], brss[[i]])
  }
  return(idx)
}

```

Now to split the training set into the predictor space and response:

```

p_space <- train_set[,1:13] #data frame of the predictor space (excluding MEDV response)
response <- train_set$MEDV #vector of response variable MEDV

model <- train_set[1:13] #select the first 13 predictors because this provides us with the highest R-sq
d = ncol(model)

M_models <- power_set(p_space, response) #stores M_1,...,M_p sets of models

```

Other helper functions to calculate variance:

```

power_var <- function(data, response_var) {
  #create empty lists sets and each_subset, which provides the RSS of every model of every list of list
  sets <- list()
  each_subset <- list()
}

```

```

new_vec <- data.frame(rep(NA, nrow(data)))
p <- ncol(data)

powers <- subsets(p)

for (i in 1:length(powers)) {
  for (j in 1:ncol(powers[[i]]) ) {
    for (k in 1:length(powers[[i]][,j]) ) {
      t = powers[[i]][k, j]
      vec <- data.frame(data[t])
      new_vec <- cbind(new_vec, vec)
    }
    new_vec <- new_vec[-1] #eliminates NA list
    new_vec <- cbind(response_var, new_vec)
    models <- (summary(lm(response_var ~ ., data = new_vec))$sigma)**2
    each_subset[[j]] <- models #creates a list of j lists for each k model
  }
  sets[[i]] <- each_subset
}
sets <- na.omit(sets)
return(sets)
}

```

Mallow's Cp, AIC, BIC, Adjusted R-squared, and CV statistics:

```

#Mallow's Cp
Cp <- function(train, response) {
  Cp_est <- rep(NA, ncol(train))
  varf <- var_fwd(train, response)
  for (i in 1:ncol(train)) {
    rsq_idx <- index_fwd(train, response)[[i]]
    var_i <- as.numeric(var_fwd(train, response)[[i]][rsq_idx])
    Cp_est[i] <- (power_rss(train, response)[[i]] + 2*ncol(train)*(var_i)/nrow(train))
  }
  return(Cp_est)
}

#AIC
AIC <- function(train, response) {
  AIC_est <- rep(NA, ncol(train))
  varf <- var_fwd(train, response)
  for (i in 1:ncol(train)) {
    rsq_idx <- index_fwd(train, response)[[i]]
    var_i <- as.numeric(var_fwd(train, response)[[i]][rsq_idx])
    AIC_est[i] <- (RSS_fwd(train, response)[[i]] + 2*ncol(train)*(var_i)/(nrow(train) *var_i ))
  }
  return(AIC_est)
}

```



```

#BIC
BIC <- function(train, response) {
  BIC_est <- rep(NA, ncol(train))
  varf <- var_fwd(train, response)
  n <- nrow(train)
  for (i in 1:ncol(train)) {
    rsq_idx <- index_fwd(train, response)[[i]]
    var_i <- as.numeric(var_fwd(train, response)[[i]][rsq_idx])
    BIC_est[i] <- (RSS_fwd(train, response)[[i]] + 2*(log(n))*ncol(train)*(var_i)/n)
  }
  return(BIC_est)
}

#Adjusted R-squared
adjR <- function(predictor_space, response_var) {#helper function that creates a list of M_k subsets for
  p = ncol(predictor_space)
  fwd_list <- list()
  index_adj <- index_fwd(predictor_space, response_var)
  adj <- rep(NA, ncol(predictor_space))

  for (k in 1:p) {
    M_set <- data.frame(predictor_space[1:k], response_var)
    rsq <- apply(M_set, 2, function(x) summary(lm(response_var ~ x))$adj.r.squared)
    names(d)[which.max(rsq)]
    idx_adj <- index_adj[[k]]

    fwd_list[[k]] <- rsq[[idx_adj]]
  }
  for (j in 1:length(fwd_list)) {
    adj[j] <- fwd_list[[j]]
  }
  return(adj)
}

#k-fold Cross-Validation
kfold <- function(predictor_space, response, k) {
  p = ncol(predictor_space)
  folds <- cut(seq(1,nrow(predictor_space)),breaks=k,labels=FALSE)
  MSE <- list()

  for (i in 1:k) {
    idx_test <- which(folds==i,arr.ind=TRUE)
    val_set <- predictor_space[idx_test, ]
    trained_set <- predictor_space[-idx_test, ]
    y_val <- response[-idx_test]

    lm_new <- lm(y_val ~ ., data = trained_set)
    pred <- predict(lm_new, newData = val_set )
    MSE <- mean((y_val - pred)^2)
  }
  return(MSE)
}

```

```

cv_fwd <- function(predictor_space, response_var, folds) {#helper function that creates a list of M_k s
  p = ncol(predictor_space)
  fwd_list <- list()
  cv <- rep(NA, length(predictor_space))

  for (k in 1:p) {
    M_set <- data.frame(predictor_space[1:k], response_var)
    fwd_list[[k]] <- kfold(M_set, response_var, folds)
  }
  for (j in 1:length(fwd_list)) {
    cv[j] <- fwd_list[[j]]
    if (cv[j] == 0) {
      cv[j] <- NA
    }
  }
  return(cv)
}

```

4. Implement Algorithm 6.2 and report the best model under C_p , BIC, adjusted R^2 and Cross-Validation (k-fold, k of your choice).

Algorithm 5.2 is the *forward stepwise selection* algorithm, which is a more computationally efficient model selection/subset selection method as opposed to best subset selection. The algorithm is as follows:

- (a) Let M_0 denote the *null* model, which contains no predictors
- (b) For $k = 0, \dots, p - 1$:
 - i. Consider all $p - k$ models that augment the predictors in M_k with one additional predictor
 - ii. Choose the best among these $p - k$ models, and call it M_{k+1} . Here *best* is defined as having smallest RSS or highest R^2 .
- (c) Select a single best model from among M_0, \dots, M_p using cross-validated prediction error, C_p , AIC, BIC, or adjusted R^2 .
- (d)

```

M_0 <- mean(response) #null model, 0 predictors, and the RSS is simply the mean of the response variabl

```

- (b)

```

index <- function(vec, value) {#indexes a set
  for (i in 1:length(vec)) {
    if (value == vec[i]) {
      yes <- i
    }
  }
  return(yes)
}

fwd <- function(predictor_space, response_var) {#helper function that creates a list of M_k subsets for
  p = ncol(predictor_space)

```

```

fwd_list <- list()

for (k in 1:p) {
  M_set <- data.frame(predictor_space[1:k], response_var)
  rsq <- apply(M_set, 2, function(x) summary(lm(response_var ~ x))$r.squared)
names(d)[which.max(rsq)]
  fwd_list[[k]] <- max(rsq[1:k])
}
return(fwd_list)
}

index_fwd <- function(predictor_space, response_var) {
  p = ncol(predictor_space)

  idx <- list()
  max_rsqa <- fwd(predictor_space, response_var)

  for (k in 1:p) {
    M_set <- data.frame(predictor_space[1:k], response_var)
    rsq <- apply(M_set, 2, function(x) summary(lm(response_var ~ x))$r.squared)
names(d)[which.max(rsq)]

    idx[[k]] <- index(rsq, max_rsqa[[k]])
  }
  return(idx)
}

fwd(p_space, response)

```

```

## [[1]]
## [1] 0.0559301
##
## [[2]]
## [1] 0.157811
##
## [[3]]
## [1] 0.157811
##
## [[4]]
## [1] 0.2369993
##
## [[5]]
## [1] 0.2369993
##
## [[6]]
## [1] 0.2369993
##
## [[7]]
## [1] 0.4772053
##
## [[8]]
## [1] 0.4772053

```

```
##
## [[9]]
## [1] 0.4772053
##
## [[10]]
## [1] 0.4772053
##
## [[11]]
## [1] 0.4772053
##
## [[12]]
## [1] 0.4772053
##
## [[13]]
## [1] 0.5457711
```

```
index_fwd(p_space, response)
```

```
## [[1]]
## [1] 1
##
## [[2]]
## [1] 2
##
## [[3]]
## [1] 2
##
## [[4]]
## [1] 4
##
## [[5]]
## [1] 4
##
## [[6]]
## [1] 4
##
## [[7]]
## [1] 7
##
## [[8]]
## [1] 7
##
## [[9]]
## [1] 7
##
## [[10]]
## [1] 7
##
## [[11]]
## [1] 7
##
## [[12]]
## [1] 7
##
```

```
## [[13]]
## [1] 13
```

So M_13 gives the highest R-squared of all the subsets. We then find the best model within these 13 predictors:

(c)

```
model <- train_set[1:13] #select the first 13 predictors because this provides us with the highest R-sq
model_RSS <- summary(fwd(p_space, train_set$MEDV)[[13]])[[6]]
d = ncol(model)

#calculate variance
var_fwd <- function(predictor_space, response_var) {#calculates variance
  p = ncol(predictor_space)
  fwd_list <- list()

  for (k in 1:p) {
    M_set <- data.frame(predictor_space[1:k], response_var)
    varfwd <- apply(M_set, 2, function(x) (summary(lm(response_var ~ x))$sigma)**2)
    names(M_set)[which.max(varfwd)]
    fwd_list[[k]] <- varfwd[1:k]
  }
  return(fwd_list)
}

#calculate min RSS for each subset
RSS_fwd <- function(predictor_space, response_var) {#calculates variance
  p = ncol(predictor_space)
  fwd_list <- list()

  for (k in 1:p) {
    M_set <- data.frame(predictor_space[1:k], response_var)
    RSSfwd <- apply(M_set, 2, function(x) (summary(lm(response_var ~ x))[[6]]))
    names(M_set)[which.max(RSSfwd)]
    fwd_list[[k]] <- min(RSSfwd[1:k])
  }
  return(fwd_list)
}
```

Thus, the 13th model within the 13th subset proves to be the one with the lowest Cp

```
#Mallow's Cp
Cp <- function(train, response) {
  Cp_est <- rep(NA, ncol(train))
  varf <- var_fwd(train, response)
  for (i in 1:ncol(train)) {
    rsq_idx <- index_fwd(train, response)[[i]]
    var_i <- as.numeric(var_fwd(train, response)[[i]][rsq_idx])
  }
}
```

```

    Cp_est[i] <- (RSS_fwd(train, response)[[i]] + 2*ncol(train)*(var_i)/nrow(train))
  }
  return(Cp_est)
}

#AIC
AIC <- function(train, response) {
  AIC_est <- rep(NA, ncol(train))
  varf <- var_fwd(train, response)
  for (i in 1:ncol(train)) {
    rsq_idx <- index_fwd(train, response)[[i]]
    var_i <- as.numeric(var_fwd(train, response)[[i]][rsq_idx])
    AIC_est[i] <- (RSS_fwd(train, response)[[i]] + 2*ncol(train)*(var_i)/(nrow(train) *var_i ))
  }
  AIC_est
}

#BIC
BIC <- function(train, response) {
  BIC_est <- rep(NA, ncol(train))
  varf <- var_fwd(train, response)
  n <- nrow(train)
  for (i in 1:ncol(train)) {
    rsq_idx <- index_fwd(train, response)[[i]]
    var_i <- as.numeric(var_fwd(train, response)[[i]][rsq_idx])
    BIC_est[i] <- (RSS_fwd(train, response)[[i]] + 2*(log(n))*ncol(train)*(var_i)/n)
  }
  BIC_est
}

#Adjusted R-squared
adjR <- function(predictor_space, response_var) {#helper function that creates a list of M_k subsets for
  p = ncol(predictor_space)
  fwd_list <- list()
  index_adj <- index_fwd(predictor_space, response_var)
  adj <- rep(NA, ncol(predictor_space))

  for (k in 1:p) {
    M_set <- data.frame(predictor_space[1:k], response_var)
    rsq <- apply(M_set, 2, function(x) summary(lm(response_var ~ x))$adj.r.squared)
names(d)[which.max(rsq)]
    idx_adj <- index_adj[[k]]

    fwd_list[[k]] <- rsq[[idx_adj]]
  }
  for (j in 1:length(fwd_list)) {
    adj[j] <- fwd_list[[j]]
  }
  adj
}

```

#k-fold Cross-Validation

```
kfold <- function(predictor_space, response, k) {  
  p = ncol(predictor_space)  
  folds <- cut(seq(1,nrow(predictor_space)),breaks=k,labels=FALSE)  
  MSE <- list()  
  
  for (i in 1:k) {  
    idx_test <- which(folds==i,arr.ind=TRUE)  
    val_set <- predictor_space[idx_test, ]  
    trained_set <- predictor_space[-idx_test, ]  
    y_val <- response[-idx_test]  
  
    lm_new <- lm(y_val ~ ., data = trained_set)  
    pred <- predict(lm_new, newData = val_set )  
    MSE <- mean((y_val - pred)^2)  
  }  
  MSE  
}
```

```
cv_fwd <- function(predictor_space, response_var, folds) {#helper function that creates a list of M_k s  
  p = ncol(predictor_space)  
  fwd_list <- list()  
  cv <- rep(NA, length(predictor_space))  
  
  for (k in 1:p) {  
    M_set <- data.frame(predictor_space[1:k], response_var)  
    fwd_list[[k]] <- kfold(M_set, response_var, folds)  
  }  
  for (j in 1:length(fwd_list)) {  
    cv[j] <- fwd_list[[j]]  
    if (cv[j] == 0) {  
      cv[j] <- NA  
    }  
  }  
  cv  
}
```

```
c <- Cp(model, response)
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
min(c)
```

```
## [1] 8.474025
```

```
index(c, min(c))
```

```
## [1] 13
```

```
aic <- AIC(model, response)
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```


[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
min(aic)
```

```
## [1] 6.302407
```

```
index(aic, min(aic))
```

```
## [1] 13
```

```
bic <- BIC(model, response)
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
na.omit(bic)
```

```
## [1] 37.35440 33.79519 33.79519 31.00746 31.00746 31.00746 22.39981  
## [8] 22.39981 22.39981 22.39981 22.39981 22.39981 19.88595
```

```
min(bic)
```

```
## [1] 19.88595
```

```
index(bic, min(bic))
```

```
## [1] 13
```

```
adj <- adjR(model, response)
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:  
## summary may be unreliable
```

[illegible]

```
## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable

## Warning in summary.lm(lm(response_var ~ x)): essentially perfect fit:
## summary may be unreliable
```

```
max(adj)
```

```
## [1] 0.5447684
```

```
index(adj, max(adj))
```

```
## [1] 13
```

```
cv <- cv_fwd(model, response, 5) #5-fold CV
cv
```

```
## [1] 8.556107e-30 1.527876e-29 1.282766e-29 8.443412e-30 1.179174e-29
## [6] 1.544564e-29 6.982719e-30 1.775587e-29 2.608442e-29 1.443355e-30
## [11] 6.315222e-30 NA 7.182102e-30
```

According to our AIC, BIC, and Adjusted-R-squared estimates, the 13th set has the best model, so we consider the linear model of all 13 predictors to be the best fit. CV states that the 11th model is the most accurate:

```
best_fwd <- lm(train_set$MEDV ~ ., data = train_set)
best_fwd
```

```
##
## Call:
```

```
## lm(formula = train_set$MEDV ~ ., data = train_set)
##
## Coefficients:
## (Intercept)          X          CRIM          ZN          INDUS
##  44.687491   -0.001749   -0.121028    0.049554    0.008859
##      CHAS          NOX          RM          AGE          DIS
##  2.455821  -20.072780    3.507981   -0.001470   -1.587636
##      RAD          TAX      PTRATIO      LSTAT
##  0.319424   -0.012316   -0.968050   -0.557563
```

5. Find the best model under LASSO and Ridge-regularized LS. Use cross-validation to choose the best penalty. You may use `glmnet` or any other library for this task.

```
grid <- 10^seq(10, -2, length = 100)
```

```
#ridge
```

```
mod_r <- cv.glmnet(as.matrix(train_set[,-14]), response, lambda = grid)
mod_r$lambda.min
```

```
## [1] 0.0231013
```

```
rstats <- cv.glmnet(as.matrix(train_set[,-14]), response, alpha = 0)
ridge_lambda <- rstats$lambda.min
```

```
ridge <- glmnet( as.matrix(train_set[,-14]), response, family = "gaussian", alpha = 0, lambda = ridge_lambda)
```

```
#lasso
```

```
mod_l <- cv.glmnet(as.matrix(train_set[,-14]), response, lambda = grid)
mod_l$lambda.min
```

```
## [1] 0.01747528
```

```
lstats <- cv.glmnet(as.matrix(train_set[,-14]), response, alpha = 1)
lasso_lambda <- lstats$lambda.min
```

```
lasso <- glmnet(as.matrix(train_set[,-14]), response, family = "gaussian", alpha = 1, lambda = lasso_lambda)
```

6. Use your 3 best models chose from the last 3 tasks to predict the housing values in your test set and compute the predicted MSE for each. Interpret your results.

```
MSE_4 <- mean((test_set$MEDV - (predict(best_fwd, test_set, type = "response")))**2)
MSE_ridge <- mean((test_set$MEDV - (predict(ridge, as.matrix(test_set[,-14]), type = "link")))**2)
MSE_lasso <- mean((test_set$MEDV - (predict(lasso, as.matrix(test_set[,-14]), type = "link")))**2)

MSE_4
```

```
## [1] 21.97067
```

```
MSE_ridge
```

```
## [1] 21.24401
```

```
MSE_lasso
```

```
## [1] 21.78773
```

Thus, the lowest MSE comes from the forward stepwise selection we did in Task 4. This is due to the fact that the least squares fit is most likely already at a low bias, and that when we use ridge, we are increasing the bias for the sake of a much lower variance. since the forward selection method has a lower MSE, then the effect of the bias must outweigh the effect of the variance. On the other hand, lasso performs better than ridge, because it is a feature selection method, so many values that would have otherwise increased bias or variance have been eliminated.

BONUS: Compare your results to the paper referenced before.