# Exercise 2

for Advanced Methods for Regression and Classification

Dzhamilia Kulikieva

31.10.2024

## Data investigation and preparation

This data set contains information about construction costs (df$y) of real estate single-family residential apartments in Tehran, Iran. Let's see the structure of the data and its variables format.

```
load("/Users/djem/Downloads/building.RData")
head(df)
```

```
##          y START.YEAR START.QUARTER COMPLETION.YEAR COMPLETION.QUARTER PhysFin1
## 1 7.696213         81             1              85                  1        1
## 2 8.517193         84             1              89                  4        1
## 3 7.090077         78             1              81                  4        1
## 4 5.105945         72             2              73                  2        1
## 5 8.612503         87             1              90                  2        1
## 6 8.556414         87             1              90                  1        1
##   PhysFin2 PhysFin3 PhysFin4 PhysFin5 PhysFin6 PhysFin7 PhysFin8   Econ1 Econ2
## 1     3150      920    598.5      190  1010.84       16     1200 6713.00  56.2
## 2     7600     1140   3040.0      400   963.81       23     2900 3152.00 106.0
## 3     4800      840    480.0      100   689.84       15      630 1627.00  41.0
## 4      685      202     13.7       20   459.54        4      140 2580.93  12.1
## 5     3000      800   1230.0      410   631.91       13     5000 6790.00 203.8
## 6     2500      640   1050.0      420   647.32       12     4800 6790.00 203.8
##    Econ3 Econ4      Econ5   Econ6 Econ7  Econ8   Econ9 Econ10 Econ11 Econ12
## 1  61.52  6.11  320957.30  3485.8  64.5 239.50 12456.6     15  797.3  809.8
## 2 103.03  3.15  685697.50  3526.1 105.5 208.80 17584.3     15 1408.4 1473.5
## 3  41.25  1.74  160401.50  1217.5  34.4 285.80  6489.1     15  614.0  608.2
## 4  10.03  1.24   38193.64   287.2  13.6  17.03   154.4     12  183.6  211.1
## 5 162.84  6.46 1640293.00 10855.3 229.3 393.30 69444.8     11 2738.8 3148.0
## 6 162.84  6.46 1640293.00 10855.3 229.3 393.30 69444.8     11 2738.8 3148.0
##    Econ13 Econ14 Econ15 Econ16   Econ17   Econ18    Econ19 Econ1.lag1
## 1 1755.00   8003  67.81  63.25  3758.77 42587.00  628132.9       4986
## 2 8842.18   8864 105.52 105.32 12113.01 45966.00 1188995.8       2700
## 3 1755.00   7773  45.91  38.34  1537.96 39066.00  524764.8       1580
## 4 1612.95   1649  11.62  10.06   392.96  8435.75  141542.6       2952
## 5 9248.40   9380 158.63 169.50 10082.00 49572.00 2318397.0       6370
## 6 9248.40   9380 158.63 169.50 10082.00 49572.00 2318397.0       6370
##   Econ2.lag1 Econ3.lag1 Econ4.lag1 Econ5.lag1 Econ6.lag1 Econ7.lag1 Econ8.lag1
## 1       55.5      60.78       3.94   297210.1     3663.5      61.50     179.63
## 2      103.0     101.84       2.65   625829.2     4386.9     100.40     156.60
## 3       40.3      40.84       1.15   150266.8     1149.5      34.10     214.35
## 4       11.6       8.50       1.99    35859.4      322.5      12.67      56.60
## 5      190.3     154.36       5.33  1523166.6    12930.0     210.70     294.98
```

```
## 6      190.3     154.36       5.33  1523166.6    12930.0     210.70     294.98
##   Econ9.lag1 Econ10.lag1 Econ11.lag1 Econ12.lag1 Econ13.lag1 Econ14.lag1
## 1    9342.45          15    757.8000      861.80     1755.00        8018
## 2   13188.23          15   1424.1000     1584.30     8776.71        8799
## 3    4866.83          15    573.7265      680.29     1755.00        6714
## 4     610.40          12    165.1000      208.60     1504.36        1582
## 5   52083.60          11   2595.2000     3000.00     9329.64        9396
## 6   52083.60          11   2595.2000     3000.00     9329.64        9396
##   Econ15.lag1 Econ16.lag1 Econ17.lag1 Econ18.lag1 Econ19.lag1 Econ1.lag2
## 1       65.00       60.53     3538.71    31940.25    610502.7       6788
## 2      101.00      101.89    13571.80    34474.50   1067772.0       3561
## 3       43.40       36.45     1535.16    29299.50    466212.2       2628
## 4       10.86        9.79      435.10    32776.00    129102.4       2649
## 5      148.76      159.00     9700.00    37179.00   1908975.7       5909
## 6      148.76      159.00     9700.00    37179.00   1908975.7       5909
##   Econ2.lag2 Econ3.lag2 Econ4.lag2 Econ5.lag2 Econ6.lag2 Econ7.lag2 Econ8.lag2
## 1       54.2      59.40       5.41   280451.7     3755.8      58.10     119.75
## 2       98.2      98.64       2.76   602224.7     3819.0      97.20     104.40
## 3       39.3      40.21       1.52   143737.7     1284.5      33.50     142.90
## 4       11.4       6.97       2.25    32793.7      388.9      11.73      42.45
## 5      177.6     147.44       6.88  1451175.9     8146.1     188.90     196.65
## 6      177.6     147.44       6.88  1451175.9     8146.1     188.90     196.65
##   Econ9.lag2 Econ10.lag2 Econ11.lag2 Econ12.lag2 Econ13.lag2 Econ14.lag2
## 1    6228.30          15     795.000      818.50     1755.00        8001
## 2    8792.15          15    1298.800     1389.60     8699.73        8735
## 3    3244.55          15     554.082      663.97     1755.00        5827
## 4     457.80          12     167.900      209.60     1450.47        1507
## 5   34722.40          11    2284.400     2627.50     9297.06        9347
## 6   34722.40          11    2284.400     2627.50     9297.06        9347
##   Econ15.lag2 Econ16.lag2 Econ17.lag2 Econ18.lag2 Econ19.lag2 Econ1.lag3
## 1       63.69       58.55     3347.72     21293.5    589389.6       5728
## 2       98.12       98.45    13596.37     22983.0    973523.7       3157
## 3       41.79       34.76     1527.55     19533.0    409677.9       2374
## 4       10.17        9.35      508.64     24582.0    123618.0       2312
## 5      140.90      146.20    10149.00     24786.0   1681849.3       7045
## 6      140.90      146.20    10149.00     24786.0   1681849.3       7045
##   Econ2.lag3 Econ3.lag3 Econ4.lag3 Econ5.lag3 Econ6.lag3 Econ7.lag3 Econ8.lag3
## 1       52.4      57.65       5.40  262789.00     2931.4      54.20       59.88
## 2       92.8      96.49       3.05  552124.40     3896.7      96.90       52.20
## 3       38.0      39.43       0.92  134548.40     1191.1      33.70       71.45
## 4       10.6       5.44       2.58   30012.46      345.3      10.79       28.30
## 5      160.0     141.34       4.72 1341072.80     8245.0     173.80       98.33
## 6      160.0     141.34       4.72 1341072.80     8245.0     173.80       98.33
##   Econ9.lag3 Econ10.lag3 Econ11.lag3 Econ12.lag3 Econ13.lag3 Econ14.lag3
## 1    3114.15          15    746.8000      815.50     1755.00        8013
## 2    4396.08          15   1294.2000     1288.00     8555.54        8585
## 3    1622.28          15    574.6000      680.50     1755.00        5565
## 4     305.20          12    180.3715      158.45     1439.00        1450
## 5   17361.20          11   2451.2000     2526.40     9254.28        9306
## 6   17361.20          11   2451.2000     2526.40     9254.28        9306
##   Econ15.lag3 Econ16.lag3 Econ17.lag3 Econ18.lag3 Econ19.lag3 Econ1.lag4
## 1       62.78       56.45     3387.72    10646.75    606524.2       7196
## 2       95.35       94.34    12063.50    11491.50    954628.6       3678
## 3       41.03       33.37     1601.79     9766.50    403875.0       2693
```

```
## 4        9.91        8.85      590.64    16388.00    121857.2       1381
## 5      136.56      138.80     9291.00    12393.00   1732937.5       5606
## 6      136.56      138.80     9291.00    12393.00   1732937.5       5606
##   Econ2.lag4 Econ3.lag4 Econ4.lag4 Econ5.lag4 Econ6.lag4 Econ7.lag4 Econ8.lag4
## 1       51.3      56.13       5.97  249110.70     2562.3      52.80     217.00
## 2       86.2      83.21       3.25  526596.40     2790.6      94.10     334.80
## 3       36.2      37.64       1.55  134312.50     1529.0      31.43     175.70
## 4       10.0       3.91       3.00   27231.21      316.5       9.85      14.15
## 5      149.1     134.80       4.09 1284199.40     6622.5     147.60     432.40
## 6      149.1     134.80       4.09 1284199.40     6622.5     147.60     432.40
##   Econ9.lag4 Econ10.lag4 Econ11.lag4 Econ12.lag4 Econ13.lag4 Econ14.lag4
## 1    10445.6          15    733.8000      815.50     1755.00        8002
## 2    14488.6          15   1143.8000     1316.30     8364.78        8393
## 3     3994.7          15    589.5000      765.80     1755.00        4930
## 4      152.6          12    197.6796      152.25     1442.31        1456
## 5    73143.5          14   2220.6000     2244.10     9231.76        9286
## 6    73143.5          14   2220.6000     2244.10     9231.76        9286
##   Econ15.lag4 Econ16.lag4 Econ17.lag4 Econ18.lag4 Econ19.lag4
## 1       60.74       54.26     2978.26       41407    601988.1
## 2       90.95       89.79    11379.37       44835    929027.1
## 3       38.70       32.04     1653.06       37933    377828.6
## 4        9.73        8.34      686.16        8194    122031.7
## 5      136.60      140.20     9821.00       48260   1734973.5
## 6      136.60      140.20     9821.00       48260   1734973.5
```

```
str(df)
```

```
## 'data.frame':    372 obs. of  108 variables:
##  $ y                 : num  7.7 8.52 7.09 5.11 8.61 ...
##  $ START.YEAR        : num  81 84 78 72 87 87 87 88 76 80 ...
##  $ START.QUARTER     : num  1 1 1 2 1 1 2 1 3 1 ...
##  $ COMPLETION.YEAR   : num  85 89 81 73 90 90 90 89 77 80 ...
##  $ COMPLETION.QUARTER: num  1 4 4 2 2 1 1 3 4 4 ...
##  $ PhysFin1          : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ PhysFin2          : num  3150 7600 4800 685 3000 2500 1810 1150 2110 3030 ...
##  $ PhysFin3          : num  920 1140 840 202 800 640 492 380 540 930 ...
##  $ PhysFin4          : num  598.5 3040 480 13.7 1230 ...
##  $ PhysFin5          : num  190 400 100 20 410 420 640 500 90 170 ...
##  $ PhysFin6          : num  1011 964 690 460 632 ...
##  $ PhysFin7          : num  16 23 15 4 13 12 11 6 5 3 ...
##  $ PhysFin8          : num  1200 2900 630 140 5000 4800 5700 5300 690 1500 ...
##  $ Econ1             : num  6713 3152 1627 2581 6790 ...
##  $ Econ2             : num  56.2 106 41 12.1 203.8 ...
##  $ Econ3             : num  61.5 103 41.2 10 162.8 ...
##  $ Econ4             : num  6.11 3.15 1.74 1.24 6.46 6.46 6.73 3.44 2.28 5.97 ...
##  $ Econ5             : num  320957 685698 160402 38194 1640293 ...
##  $ Econ6             : num  3486 3526 1218 287 10855 ...
##  $ Econ7             : num  64.5 105.5 34.4 13.6 229.3 ...
##  $ Econ8             : num  240 209 286 17 393 ...
##  $ Econ9             : num  12457 17584 6489 154 69445 ...
##  $ Econ10            : int  15 15 15 12 11 11 11 11 15 15 ...
##  $ Econ11            : num  797 1408 614 184 2739 ...
##  $ Econ12            : num  810 1474 608 211 3148 ...
##  $ Econ13            : num  1755 8842 1755 1613 9248 ...
##  $ Econ14            : num  8003 8864 7773 1649 9380 ...
```

```
##  $ Econ15           : num  67.8 105.5 45.9 11.6 158.6 ...
##  $ Econ16           : num  63.2 105.3 38.3 10.1 169.5 ...
##  $ Econ17           : num  3759 12113 1538 393 10082 ...
##  $ Econ18           : num  42587 45966 39066 8436 49572 ...
##  $ Econ19           : num  628133 1188996 524765 141543 2318397 ...
##  $ Econ1.lag1       : num  4986 2700 1580 2952 6370 ...
##  $ Econ2.lag1       : num  55.5 103 40.3 11.6 190.3 ...
##  $ Econ3.lag1       : num  60.8 101.8 40.8 8.5 154.4 ...
##  $ Econ4.lag1       : num  3.94 2.65 1.15 1.99 5.33 5.33 6.46 3.8 2.32 4.3 ...
##  $ Econ5.lag1       : num  297210 625829 150267 35859 1523167 ...
##  $ Econ6.lag1       : num  3664 4387 1150 322 12930 ...
##  $ Econ7.lag1       : num  61.5 100.4 34.1 12.7 210.7 ...
##  $ Econ8.lag1       : num  179.6 156.6 214.3 56.6 295 ...
##  $ Econ9.lag1       : num  9342 13188 4867 610 52084 ...
##  $ Econ10.lag1      : int  15 15 15 12 11 11 11 11 15 15 ...
##  $ Econ11.lag1      : num  758 1424 574 165 2595 ...
##  $ Econ12.lag1      : num  862 1584 680 209 3000 ...
##  $ Econ13.lag1      : num  1755 8777 1755 1504 9330 ...
##  $ Econ14.lag1      : num  8018 8799 6714 1582 9396 ...
##  $ Econ15.lag1      : num  65 101 43.4 10.9 148.8 ...
##  $ Econ16.lag1      : num  60.53 101.89 36.45 9.79 159 ...
##  $ Econ17.lag1      : num  3539 13572 1535 435 9700 ...
##  $ Econ18.lag1      : num  31940 34474 29300 32776 37179 ...
##  $ Econ19.lag1      : num  610503 1067772 466212 129102 1908976 ...
##  $ Econ1.lag2       : num  6788 3561 2628 2649 5909 ...
##  $ Econ2.lag2       : num  54.2 98.2 39.3 11.4 177.6 ...
##  $ Econ3.lag2       : num  59.4 98.64 40.21 6.97 147.44 ...
##  $ Econ4.lag2       : num  5.41 2.76 1.52 2.25 6.88 6.88 5.33 6.54 2.69 3.53 ...
##  $ Econ5.lag2       : num  280452 602225 143738 32794 1451176 ...
##  $ Econ6.lag2       : num  3756 3819 1284 389 8146 ...
##  $ Econ7.lag2       : num  58.1 97.2 33.5 11.7 188.9 ...
##  $ Econ8.lag2       : num  119.8 104.4 142.9 42.5 196.7 ...
##  $ Econ9.lag2       : num  6228 8792 3245 458 34722 ...
##  $ Econ10.lag2      : int  15 15 15 12 11 11 11 11 15 15 ...
##  $ Econ11.lag2      : num  795 1299 554 168 2284 ...
##  $ Econ12.lag2      : num  818 1390 664 210 2628 ...
##  $ Econ13.lag2      : num  1755 8700 1755 1450 9297 ...
##  $ Econ14.lag2      : num  8001 8735 5827 1507 9347 ...
##  $ Econ15.lag2      : num  63.7 98.1 41.8 10.2 140.9 ...
##  $ Econ16.lag2      : num  58.55 98.45 34.76 9.35 146.2 ...
##  $ Econ17.lag2      : num  3348 13596 1528 509 10149 ...
##  $ Econ18.lag2      : num  21294 22983 19533 24582 24786 ...
##  $ Econ19.lag2      : num  589390 973524 409678 123618 1681849 ...
##  $ Econ1.lag3       : num  5728 3157 2374 2312 7045 ...
##  $ Econ2.lag3       : num  52.4 92.8 38 10.6 160 ...
##  $ Econ3.lag3       : num  57.65 96.49 39.43 5.44 141.34 ...
##  $ Econ4.lag3       : num  5.4 3.05 0.92 2.58 4.72 4.72 6.88 6.73 2.68 3.39 ...
##  $ Econ5.lag3       : num  262789 552124 134548 30012 1341073 ...
##  $ Econ6.lag3       : num  2931 3897 1191 345 8245 ...
##  $ Econ7.lag3       : num  54.2 96.9 33.7 10.8 173.8 ...
##  $ Econ8.lag3       : num  59.9 52.2 71.5 28.3 98.3 ...
##  $ Econ9.lag3       : num  3114 4396 1622 305 17361 ...
##  $ Econ10.lag3      : int  15 15 15 12 11 11 11 11 15 15 ...
##  $ Econ11.lag3      : num  747 1294 575 180 2451 ...
```

```
##  $ Econ12.lag3       : num  816 1288 680 158 2526 ...
##  $ Econ13.lag3       : num  1755 8556 1755 1439 9254 ...
##  $ Econ14.lag3       : num  8013 8585 5565 1450 9306 ...
##  $ Econ15.lag3       : num  62.78 95.35 41.03 9.91 136.56 ...
##  $ Econ16.lag3       : num  56.45 94.34 33.37 8.85 138.8 ...
##  $ Econ17.lag3       : num  3388 12064 1602 591 9291 ...
##  $ Econ18.lag3       : num  10647 11492 9766 16388 12393 ...
##  $ Econ19.lag3       : num  606524 954629 403875 121857 1732938 ...
##  $ Econ1.lag4        : num  7196 3678 2693 1381 5606 ...
##  $ Econ2.lag4        : num  51.3 86.2 36.2 10 149.1 ...
##  $ Econ3.lag4        : num  56.13 83.21 37.64 3.91 134.8 ...
##  $ Econ4.lag4        : num  5.97 3.25 1.55 3 4.09 4.09 4.72 6.46 3.56 3.25 ...
##  $ Econ5.lag4        : num  249111 526596 134312 27231 1284199 ...
##  $ Econ6.lag4        : num  2562 2791 1529 316 6622 ...
##  $ Econ7.lag4        : num  52.8 94.1 31.43 9.85 147.6 ...
##  $ Econ8.lag4        : num  217 334.8 175.7 14.2 432.4 ...
##  $ Econ9.lag4        : num  10446 14489 3995 153 73144 ...
##  $ Econ10.lag4       : int  15 15 15 12 14 14 11 11 15 15 ...
##    [list output truncated]
```

To begin to work with this data set, let's split the samples randomly into training (2/3) and test (1/3) data.

```
set.seed(2024)
sample_index <- sample(1:nrow(df), size = floor(2/3 * nrow(df)))
train_data <- df[sample_index, ]
test_data <- df[-sample_index, ]
```

# 1. Computing the full model with lm() using train data

```
model <- lm(y ~ ., data = train_data)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54334 -0.10668  0.00756  0.12160  0.38711
##
## Coefficients: (34 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.962e+02  2.337e+02   0.839  0.40237
## START.YEAR          -2.695e+00  3.330e+00  -0.809  0.41943
## START.QUARTER       -9.545e-02  1.002e+00  -0.095  0.92424
## COMPLETION.YEAR      5.318e-02  3.438e-02   1.547  0.12366
## COMPLETION.QUARTER   4.510e-02  1.748e-02   2.580  0.01070 *
## PhysFin1            -3.032e-02  4.342e-03  -6.983 5.86e-11 ***
## PhysFin2             4.067e-05  4.337e-05   0.938  0.34963
## PhysFin3            -1.981e-04  1.292e-04  -1.533  0.12705
## PhysFin4             4.464e-06  7.255e-05   0.062  0.95101
## PhysFin5            -3.163e-03  6.211e-04  -5.092 9.12e-07 ***
## PhysFin6             6.631e-04  1.059e-04   6.261 2.90e-09 ***
## PhysFin7                    NA         NA      NA       NA
```

```
## PhysFin8           4.687e-04  2.933e-05  15.979  < 2e-16 ***
## Econ1             -2.722e-04  4.712e-04  -0.578  0.56419
## Econ2              5.227e-01  7.732e-01   0.676  0.49995
## Econ3              5.412e-01  1.617e-01   3.346  0.00100 **
## Econ4             -2.898e-01  2.478e-01  -1.169  0.24382
## Econ5             -7.364e-05  9.251e-05  -0.796  0.42709
## Econ6             -5.831e-04  1.255e-03  -0.465  0.64282
## Econ7             -2.806e-01  1.029e-01  -2.728  0.00703 **
## Econ8              1.528e-02  7.967e-03   1.918  0.05673 .
## Econ9             -5.578e-05  1.576e-04  -0.354  0.72387
## Econ10             5.324e-02  5.283e-01   0.101  0.91985
## Econ11             8.483e-04  2.331e-03   0.364  0.71632
## Econ12             7.382e-05  2.752e-03   0.027  0.97863
## Econ13            -1.134e-04  1.691e-04  -0.670  0.50351
## Econ14             1.440e-04  7.605e-04   0.189  0.85002
## Econ15             1.005e-02  2.825e-01   0.036  0.97166
## Econ16             3.928e-01  3.592e-01   1.093  0.27570
## Econ17             1.031e-03  6.570e-04   1.570  0.11825
## Econ18             1.054e-04  1.089e-04   0.967  0.33482
## Econ19            -9.845e-06  1.176e-05  -0.838  0.40343
## Econ1.lag1        -5.667e-04  5.575e-04  -1.017  0.31080
## Econ2.lag1        -1.279e+00  5.125e-01  -2.495  0.01353 *
## Econ3.lag1         6.410e-02  1.686e-01   0.380  0.70432
## Econ4.lag1         7.224e-01  1.415e+00   0.511  0.61034
## Econ5.lag1         9.550e-05  4.976e-05   1.919  0.05659 .
## Econ6.lag1         9.806e-04  7.806e-04   1.256  0.21069
## Econ7.lag1        -1.901e-01  1.436e-01  -1.324  0.18724
## Econ8.lag1         3.335e-04  6.478e-03   0.051  0.95900
## Econ9.lag1        -7.041e-05  9.827e-05  -0.716  0.47466
## Econ10.lag1       -2.608e-01  2.778e-01  -0.939  0.34916
## Econ11.lag1        3.434e-03  7.408e-03   0.464  0.64354
## Econ12.lag1       -1.034e-03  2.971e-03  -0.348  0.72825
## Econ13.lag1       -5.827e-04  5.185e-04  -1.124  0.26267
## Econ14.lag1        1.592e-04  9.750e-04   0.163  0.87046
## Econ15.lag1        9.217e-01  3.960e-01   2.327  0.02111 *
## Econ16.lag1       -8.690e-02  6.677e-01  -0.130  0.89661
## Econ17.lag1       -8.737e-04  1.095e-03  -0.798  0.42598
## Econ18.lag1        5.566e-05  1.142e-04   0.487  0.62660
## Econ19.lag1       -6.118e-06  8.584e-06  -0.713  0.47697
## Econ1.lag2        -1.683e-04  1.933e-04  -0.871  0.38504
## Econ2.lag2         3.768e-01  5.463e-01   0.690  0.49126
## Econ3.lag2        -3.305e-01  3.060e-01  -1.080  0.28157
## Econ4.lag2        -9.106e-02  3.961e-01  -0.230  0.81846
## Econ5.lag2         7.188e-05  1.330e-04   0.541  0.58950
## Econ6.lag2         1.010e-03  1.029e-03   0.981  0.32802
## Econ7.lag2         1.499e-01  1.300e-01   1.153  0.25044
## Econ8.lag2        -2.471e-02  1.103e-02  -2.241  0.02629 *
## Econ9.lag2        -8.739e-06  1.083e-04  -0.081  0.93578
## Econ10.lag2        7.319e-02  4.279e-01   0.171  0.86438
## Econ11.lag2        2.114e-03  5.099e-03   0.415  0.67894
## Econ12.lag2       -8.716e-04  7.115e-03  -0.123  0.90264
## Econ13.lag2       -3.862e-04  6.125e-04  -0.631  0.52916
## Econ14.lag2       -2.228e-04  2.264e-04  -0.984  0.32634
## Econ15.lag2       -6.272e-01  3.193e-01  -1.964  0.05109 .
```

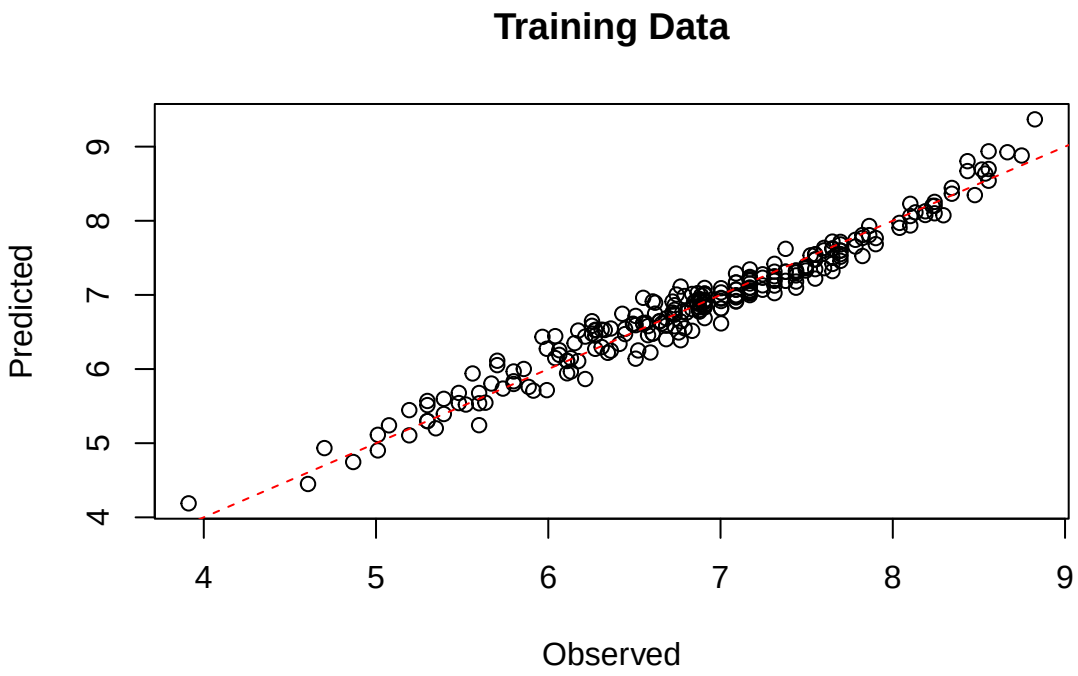
```
## Econ16.lag2        -2.809e-01  8.737e-01  -0.322  0.74818
## Econ17.lag2        -7.683e-05  7.200e-04  -0.107  0.91514
## Econ18.lag2         1.410e-04  1.111e-04   1.269  0.20622
## Econ19.lag2        -2.970e-06  3.334e-06  -0.891  0.37433
## Econ1.lag3         -3.557e-04  6.810e-04  -0.522  0.60213
## Econ2.lag3         -3.996e-01  2.834e-01  -1.410  0.16033
## Econ3.lag3          5.925e-01  4.682e-01   1.266  0.20731
## Econ4.lag3          1.578e-01  2.830e-01   0.558  0.57786
## Econ5.lag3         -3.475e-05  2.933e-05  -1.185  0.23767
## Econ6.lag3                 NA         NA      NA       NA
## Econ7.lag3                 NA         NA      NA       NA
## Econ8.lag3                 NA         NA      NA       NA
## Econ9.lag3                 NA         NA      NA       NA
## Econ10.lag3                NA         NA      NA       NA
## Econ11.lag3                NA         NA      NA       NA
## Econ12.lag3                NA         NA      NA       NA
## Econ13.lag3                NA         NA      NA       NA
## Econ14.lag3                NA         NA      NA       NA
## Econ15.lag3                NA         NA      NA       NA
## Econ16.lag3                NA         NA      NA       NA
## Econ17.lag3                NA         NA      NA       NA
## Econ18.lag3                NA         NA      NA       NA
## Econ19.lag3                NA         NA      NA       NA
## Econ1.lag4                 NA         NA      NA       NA
## Econ2.lag4                 NA         NA      NA       NA
## Econ3.lag4                 NA         NA      NA       NA
## Econ4.lag4                 NA         NA      NA       NA
## Econ5.lag4                 NA         NA      NA       NA
## Econ6.lag4                 NA         NA      NA       NA
## Econ7.lag4                 NA         NA      NA       NA
## Econ8.lag4                 NA         NA      NA       NA
## Econ9.lag4                 NA         NA      NA       NA
## Econ10.lag4                NA         NA      NA       NA
## Econ11.lag4                NA         NA      NA       NA
## Econ12.lag4                NA         NA      NA       NA
## Econ13.lag4                NA         NA      NA       NA
## Econ14.lag4                NA         NA      NA       NA
## Econ15.lag4                NA         NA      NA       NA
## Econ16.lag4                NA         NA      NA       NA
## Econ17.lag4                NA         NA      NA       NA
## Econ18.lag4                NA         NA      NA       NA
## Econ19.lag4                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.21 on 174 degrees of freedom
## Multiple R-squared:  0.9589, Adjusted R-squared:  0.9416
## F-statistic:  55.6 on 73 and 174 DF,  p-value: < 2.2e-16
```

There are a few significant predictors with p-values below 0.05, such as PhysFin1, PhysFin5, PhysFin6, Econ2.lag1, among others. The large number of variables with high p-values suggests that they may not be significant to the model and could be removed for simplification.

### 1(a) Visualisation of fitted values versus response.

```
predicted_train <- predict(model, train_data)
```

```
plot(train_data$y, predicted_train, xlab = 'Observed', ylab = 'Predicted',
main = 'Training Data')
abline(0,1, col = 'red', lty = 2)
```




The points are generally close to the line, indicating good model performance for most data. However, there are some outliers (notably in the top right) where predicted values significantly deviate, suggesting that the model struggles with extreme values or there are anomalies.

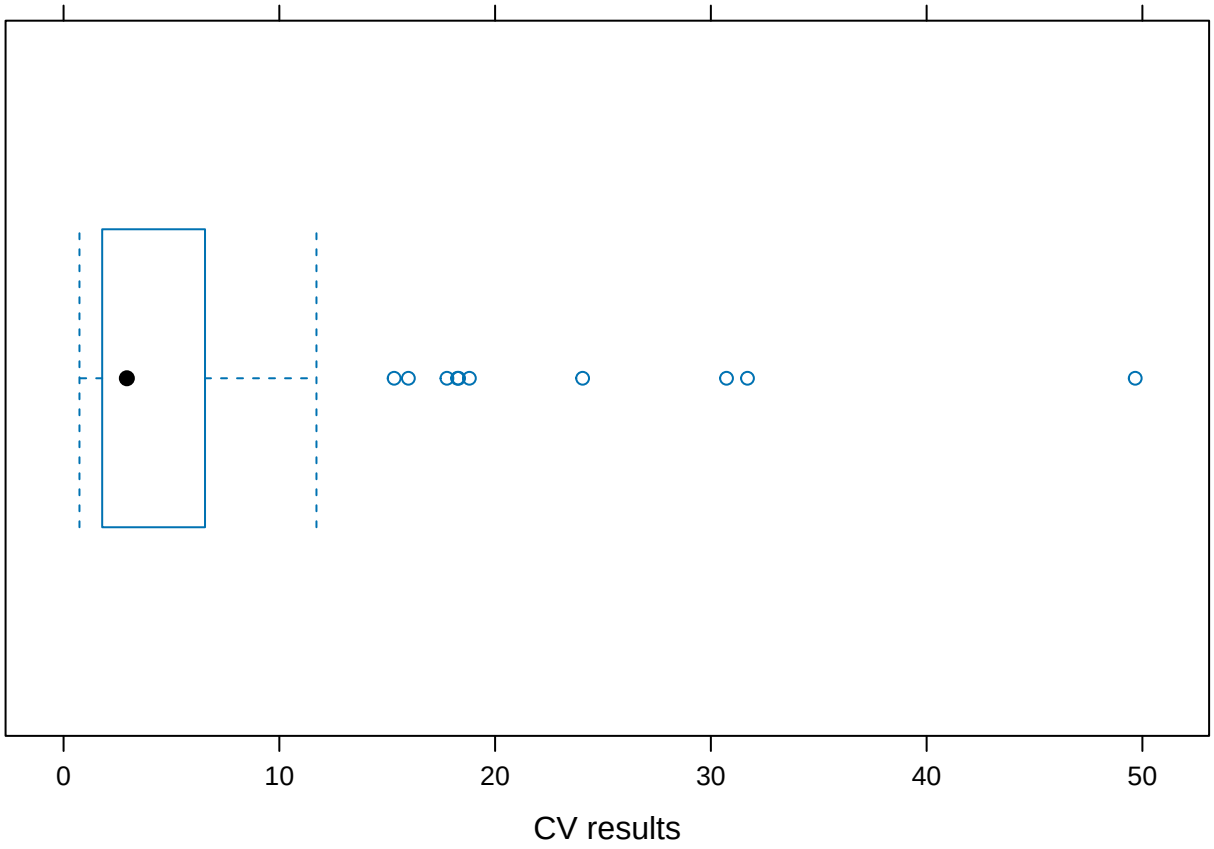
```
rmse_train <- sqrt(mean((train_data$y - predicted_train)^2))
print(rmse_train)
```

```
## [1] 0.1759123
```

With an RMSE of 0.1759, the model demonstrates a good fit, indicating small average prediction errors.

### 1(b) Using cvFit() function for model evaluation based on crossvalidation (CV).

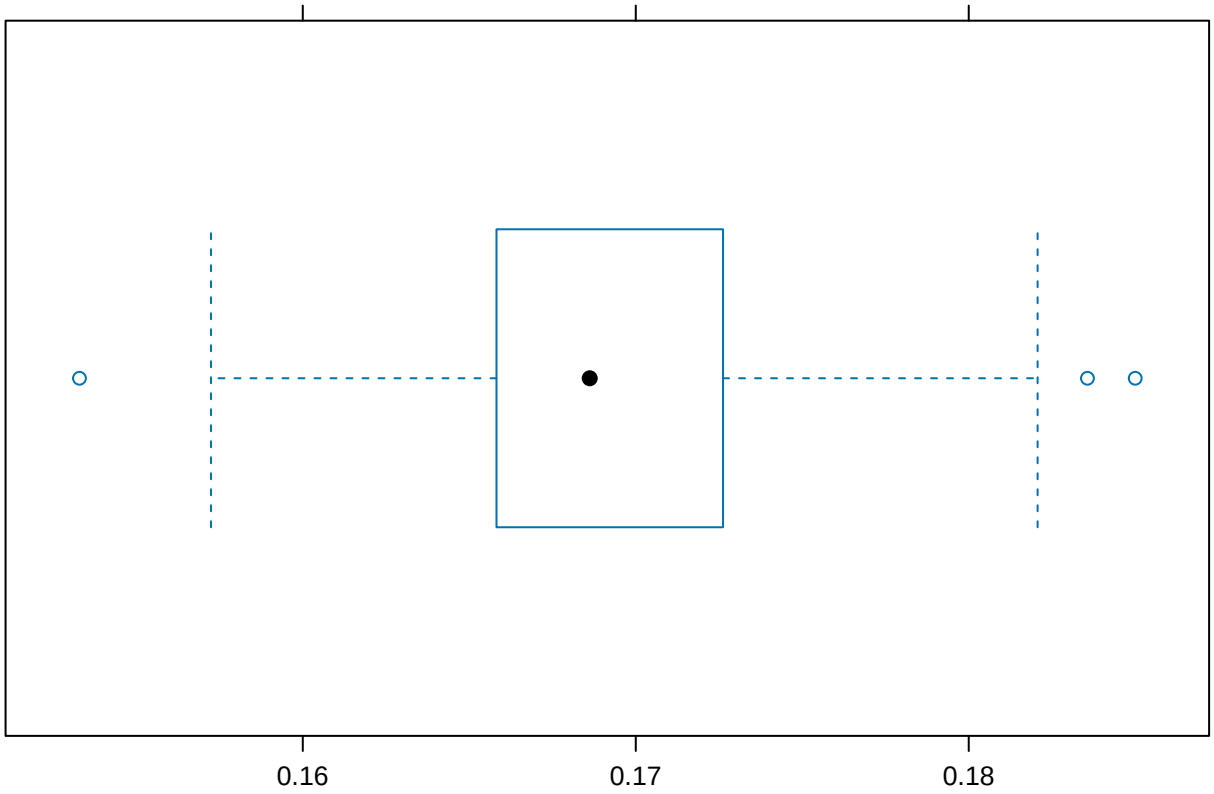
```
set.seed(123)
cv_result <- cvFit(model, data = train_data, y = train_data$y, cost = rmspe,
                   K = 5, R = 100)
plot(cv_result)
```




Most metric values fall within a narrow range (inside the box), indicating that the model shows fairly consistent results in the majority of folds. However, the presence of outliers suggests that the model performed noticeably worse in some folds.

### 1(c) Using cost=rtmspe to handle outliers.

```
set.seed(123)
cv_result_rtmspe <- cvFit(model, data = train_data, y = train_data$y, cost = rtmspe,
                          K = 5, R = 100)
plot(cv_result_rtmspe)
```




Here RTMSPE measures the percentage difference between predicted and actual values, but unlike RMSPE, it trims a certain percentage of extreme values to reduce the effect of outliers.

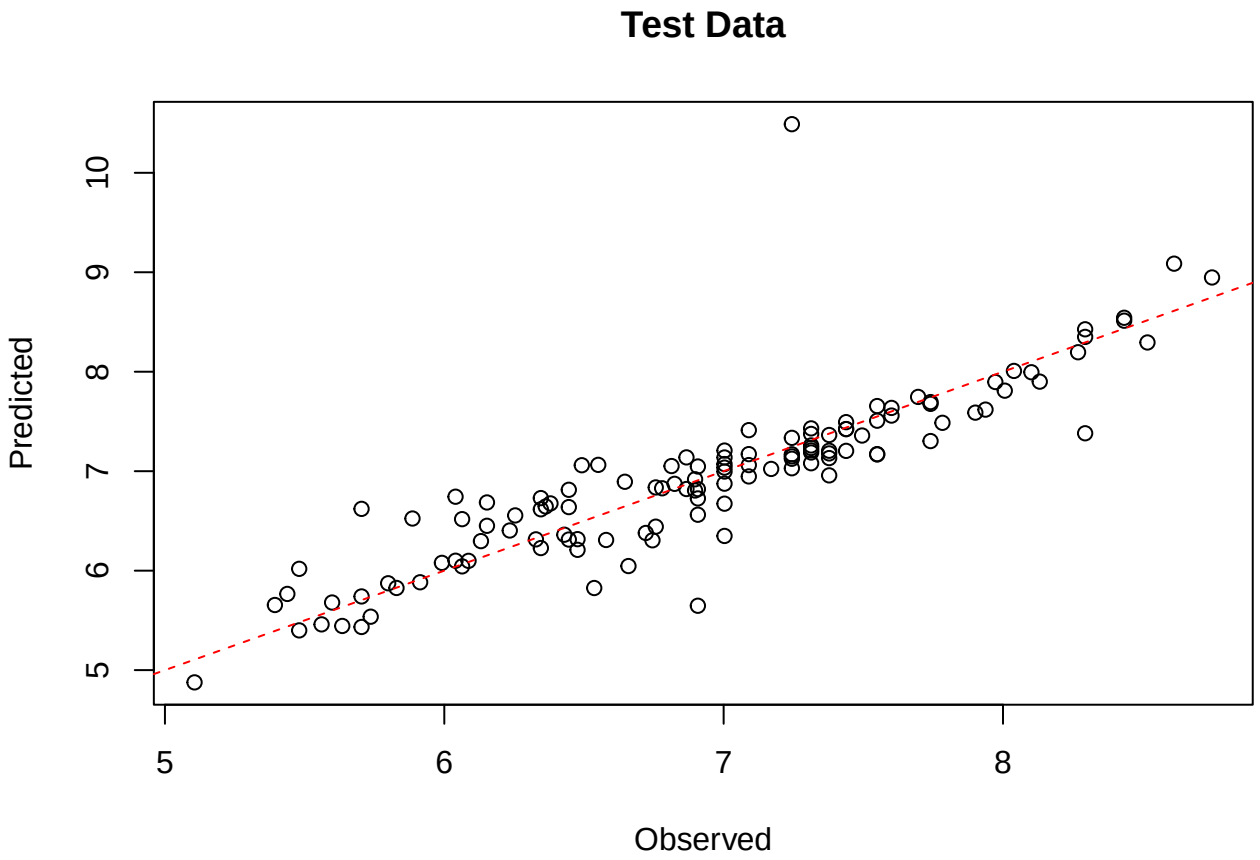
This trimming helps make the evaluation more representative of the model's general performance. As we can see, the boxplot in this case shows more stable results: the median is in the center, a narrow interquartile range and fewer outliers.

### 1(d) Predicting the response for the test data and visualizing the result.

```
predicted_test <- predict(model, test_data)

rmse_test <- sqrt(mean((test_data$y - predicted_test)^2))
print(rmse_test)
```

```
## [1] 0.4190339
```

```
plot(test_data$y, predicted_test, xlab = 'Observed', ylab = 'Predicted',
     main = 'Test Data')
abline(0, 1, col = 'red', lty = 2)
```




The model shows reasonable accuracy on the test data, but there is greater variance compared to the training data, which may indicate slight overfitting. The presence of outliers suggests the need for model improvement.

With an RMSE of 0.4190 on the test data, the model's error is larger compared to the training set (vs 0.1759 on train data), indicating reduced accuracy when applied to unseen data. This suggests that the model might not generalize as well, possibly due to overfitting.

## 2. Best subset regression with the function regsubsets().

As number of predictors in the initial model is too big, in order to use regsubsets() we need first to reduce the predictors.

### 2(a) Using stepAIC to perform step-by-step elimination of insignificant predictors, and then applying the regsubsets() function to the result.

```
reduced_model <- stepAIC(model, direction = 'backward', trace = FALSE)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = y ~ START.YEAR + COMPLETION.YEAR + COMPLETION.QUARTER +
##     PhysFin1 + PhysFin2 + PhysFin3 + PhysFin5 + PhysFin6 + PhysFin8 +
##     Econ1 + Econ2 + Econ3 + Econ4 + Econ5 + Econ6 + Econ7 + Econ8 +
##     Econ9 + Econ13 + Econ16 + Econ17 + Econ18 + Econ19 + Econ1.lag1 +
##     Econ2.lag1 + Econ3.lag1 + Econ4.lag1 + Econ5.lag1 + Econ6.lag1 +
##     Econ7.lag1 + Econ9.lag1 + Econ10.lag1 + Econ11.lag1 + Econ12.lag1 +
##     Econ13.lag1 + Econ15.lag1 + Econ17.lag1 + Econ18.lag1 + Econ19.lag1 +
```

```
##     Econ1.lag2 + Econ2.lag2 + Econ3.lag2 + Econ5.lag2 + Econ6.lag2 +
##     Econ7.lag2 + Econ8.lag2 + Econ11.lag2 + Econ13.lag2 + Econ15.lag2 +
##     Econ16.lag2 + Econ18.lag2 + Econ19.lag2 + Econ1.lag3 + Econ2.lag3 +
##     Econ3.lag3 + Econ4.lag3 + Econ5.lag3, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55810 -0.10554  0.01286  0.12368  0.39056
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.018e+02  3.657e+01   2.785 0.005897 **
## START.YEAR          -1.372e+00  5.270e-01  -2.603 0.009976 **
## COMPLETION.YEAR      6.342e-02  3.144e-02   2.017 0.045105 *
## COMPLETION.QUARTER   4.664e-02  1.610e-02   2.896 0.004221 **
## PhysFin1            -3.169e-02  3.575e-03  -8.866 5.48e-16 ***
## PhysFin2             4.193e-05  3.117e-05   1.345 0.180154
## PhysFin3            -2.029e-04  1.187e-04  -1.709 0.089129 .
## PhysFin5            -3.218e-03  5.031e-04  -6.396 1.21e-09 ***
## PhysFin6             6.624e-04  9.522e-05   6.956 5.49e-11 ***
## PhysFin8             4.683e-04  2.789e-05  16.788  < 2e-16 ***
## Econ1               -2.824e-04  7.690e-05  -3.673 0.000312 ***
## Econ2                3.373e-01  9.225e-02   3.657 0.000331 ***
## Econ3                3.801e-01  6.113e-02   6.218 3.13e-09 ***
## Econ4               -1.776e-01  9.380e-02  -1.894 0.059760 .
## Econ5               -5.330e-05  9.101e-06  -5.857 2.04e-08 ***
## Econ6               -4.451e-04  1.231e-04  -3.615 0.000385 ***
## Econ7               -1.955e-01  3.155e-02  -6.196 3.51e-09 ***
## Econ8                1.135e-02  2.146e-03   5.289 3.37e-07 ***
## Econ9               -4.418e-05  1.292e-05  -3.418 0.000772 ***
## Econ13              -9.867e-05  3.683e-05  -2.679 0.008028 **
## Econ16               2.827e-01  7.424e-02   3.808 0.000189 ***
## Econ17               6.223e-04  1.136e-04   5.476 1.37e-07 ***
## Econ18               7.300e-05  1.855e-05   3.936 0.000116 ***
## Econ19              -8.336e-06  1.665e-06  -5.006 1.27e-06 ***
## Econ1.lag1          -4.500e-04  9.182e-05  -4.901 2.04e-06 ***
## Econ2.lag1          -8.760e-01  1.850e-01  -4.735 4.27e-06 ***
## Econ3.lag1           4.026e-02  2.308e-02   1.744 0.082740 .
## Econ4.lag1           6.046e-01  1.501e-01   4.027 8.15e-05 ***
## Econ5.lag1           7.767e-05  1.499e-05   5.180 5.64e-07 ***
## Econ6.lag1           7.276e-04  1.278e-04   5.693 4.67e-08 ***
## Econ7.lag1          -1.256e-01  2.415e-02  -5.200 5.12e-07 ***
## Econ9.lag1          -5.877e-05  1.724e-05  -3.408 0.000799 ***
## Econ10.lag1         -1.822e-01  8.665e-02  -2.103 0.036806 *
## Econ11.lag1          1.695e-03  5.321e-04   3.185 0.001693 **
## Econ12.lag1         -6.526e-04  5.090e-04  -1.282 0.201368
## Econ13.lag1         -3.864e-04  1.159e-04  -3.333 0.001031 **
## Econ15.lag1          6.103e-01  1.268e-01   4.814 3.01e-06 ***
## Econ17.lag1         -5.790e-04  1.184e-04  -4.890 2.14e-06 ***
## Econ18.lag1          3.682e-05  1.722e-05   2.138 0.033817 *
## Econ19.lag1         -3.007e-06  1.232e-06  -2.440 0.015613 *
## Econ1.lag2          -2.088e-04  6.042e-05  -3.456 0.000675 ***
## Econ2.lag2           2.223e-01  9.331e-02   2.382 0.018183 *
## Econ3.lag2          -1.947e-01  4.156e-02  -4.684 5.34e-06 ***
```

```
## Econ5.lag2          4.634e-05  9.137e-06   5.072 9.32e-07 ***
## Econ6.lag2          6.159e-04  1.329e-04   4.636 6.60e-06 ***
## Econ7.lag2          1.048e-01  2.177e-02   4.814 3.01e-06 ***
## Econ8.lag2         -1.664e-02  3.197e-03  -5.207 4.97e-07 ***
## Econ11.lag2         1.116e-03  4.199e-04   2.658 0.008537 **
## Econ13.lag2        -2.868e-04  9.488e-05  -3.022 0.002854 **
## Econ15.lag2        -4.689e-01  1.134e-01  -4.136 5.31e-05 ***
## Econ16.lag2        -2.698e-01  6.716e-02  -4.017 8.48e-05 ***
## Econ18.lag2         9.468e-05  2.164e-05   4.375 2.00e-05 ***
## Econ19.lag2        -2.755e-06  1.185e-06  -2.325 0.021111 *
## Econ1.lag3         -2.824e-04  9.165e-05  -3.082 0.002365 **
## Econ2.lag3         -1.957e-01  5.479e-02  -3.572 0.000450 ***
## Econ3.lag3          3.788e-01  7.136e-02   5.308 3.08e-07 ***
## Econ4.lag3          1.151e-01  8.210e-02   1.402 0.162508
## Econ5.lag3         -2.753e-05  8.839e-06  -3.115 0.002122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2034 on 190 degrees of freedom
## Multiple R-squared:  0.9579, Adjusted R-squared:  0.9452
## F-statistic: 75.81 on 57 and 190 DF,  p-value: < 2.2e-16
```

Now we have reduced model to 57 predictors out of 190. Let's try to reduced the number of predictors to 50 (as it's recommended in the task) by applying step() function with backward selection method. steps = 7 is used to limit the number of variables in the model and reduce the model from 57 predictors to about 50.

```r
initial_model <- lm(y ~ START.YEAR + COMPLETION.YEAR + COMPLETION.QUARTER +
    PhysFin1 + PhysFin2 + PhysFin3 + PhysFin5 + PhysFin6 + PhysFin8 +
    Econ1 + Econ2 + Econ3 + Econ4 + Econ5 + Econ6 + Econ7 + Econ8 +
    Econ9 + Econ13 + Econ16 + Econ17 + Econ18 + Econ19 + Econ1.lag1 +
    Econ2.lag1 + Econ3.lag1 + Econ4.lag1 + Econ5.lag1 + Econ6.lag1 +
    Econ7.lag1 + Econ9.lag1 + Econ10.lag1 + Econ11.lag1 + Econ12.lag1 +
    Econ13.lag1 + Econ15.lag1 + Econ17.lag1 + Econ18.lag1 + Econ19.lag1 +
    Econ1.lag2 + Econ2.lag2 + Econ3.lag2 + Econ5.lag2 + Econ6.lag2 +
    Econ7.lag2 + Econ8.lag2 + Econ11.lag2 + Econ13.lag2 + Econ15.lag2 +
    Econ16.lag2 + Econ18.lag2 + Econ19.lag2 + Econ1.lag3 + Econ2.lag3 +
    Econ3.lag3 + Econ4.lag3 + Econ5.lag3, data = train_data)


final_model <- step(initial_model, direction = "backward", steps = 7, trace = FALSE)

summary(final_model)
```

```
##
## Call:
## lm(formula = y ~ START.YEAR + COMPLETION.YEAR + COMPLETION.QUARTER +
##     PhysFin1 + PhysFin2 + PhysFin3 + PhysFin5 + PhysFin6 + PhysFin8 +
##     Econ1 + Econ2 + Econ3 + Econ4 + Econ5 + Econ6 + Econ7 + Econ8 +
##     Econ9 + Econ13 + Econ16 + Econ17 + Econ18 + Econ19 + Econ1.lag1 +
##     Econ2.lag1 + Econ3.lag1 + Econ4.lag1 + Econ5.lag1 + Econ6.lag1 +
##     Econ7.lag1 + Econ9.lag1 + Econ10.lag1 + Econ11.lag1 + Econ12.lag1 +
##     Econ13.lag1 + Econ15.lag1 + Econ17.lag1 + Econ18.lag1 + Econ19.lag1 +
##     Econ1.lag2 + Econ2.lag2 + Econ3.lag2 + Econ5.lag2 + Econ6.lag2 +
##     Econ7.lag2 + Econ8.lag2 + Econ11.lag2 + Econ13.lag2 + Econ15.lag2 +
##     Econ16.lag2 + Econ18.lag2 + Econ19.lag2 + Econ1.lag3 + Econ2.lag3 +
```

```
##     Econ3.lag3 + Econ4.lag3 + Econ5.lag3, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55810 -0.10554  0.01286  0.12368  0.39056
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.018e+02  3.657e+01   2.785 0.005897 **
## START.YEAR          -1.372e+00  5.270e-01  -2.603 0.009976 **
## COMPLETION.YEAR      6.342e-02  3.144e-02   2.017 0.045105 *
## COMPLETION.QUARTER   4.664e-02  1.610e-02   2.896 0.004221 **
## PhysFin1            -3.169e-02  3.575e-03  -8.866 5.48e-16 ***
## PhysFin2             4.193e-05  3.117e-05   1.345 0.180154
## PhysFin3            -2.029e-04  1.187e-04  -1.709 0.089129 .
## PhysFin5            -3.218e-03  5.031e-04  -6.396 1.21e-09 ***
## PhysFin6             6.624e-04  9.522e-05   6.956 5.49e-11 ***
## PhysFin8             4.683e-04  2.789e-05  16.788  < 2e-16 ***
## Econ1               -2.824e-04  7.690e-05  -3.673 0.000312 ***
## Econ2                3.373e-01  9.225e-02   3.657 0.000331 ***
## Econ3                3.801e-01  6.113e-02   6.218 3.13e-09 ***
## Econ4               -1.776e-01  9.380e-02  -1.894 0.059760 .
## Econ5               -5.330e-05  9.101e-06  -5.857 2.04e-08 ***
## Econ6               -4.451e-04  1.231e-04  -3.615 0.000385 ***
## Econ7               -1.955e-01  3.155e-02  -6.196 3.51e-09 ***
## Econ8                1.135e-02  2.146e-03   5.289 3.37e-07 ***
## Econ9               -4.418e-05  1.292e-05  -3.418 0.000772 ***
## Econ13              -9.867e-05  3.683e-05  -2.679 0.008028 **
## Econ16               2.827e-01  7.424e-02   3.808 0.000189 ***
## Econ17               6.223e-04  1.136e-04   5.476 1.37e-07 ***
## Econ18               7.300e-05  1.855e-05   3.936 0.000116 ***
## Econ19              -8.336e-06  1.665e-06  -5.006 1.27e-06 ***
## Econ1.lag1          -4.500e-04  9.182e-05  -4.901 2.04e-06 ***
## Econ2.lag1          -8.760e-01  1.850e-01  -4.735 4.27e-06 ***
## Econ3.lag1           4.026e-02  2.308e-02   1.744 0.082740 .
## Econ4.lag1           6.046e-01  1.501e-01   4.027 8.15e-05 ***
## Econ5.lag1           7.767e-05  1.499e-05   5.180 5.64e-07 ***
## Econ6.lag1           7.276e-04  1.278e-04   5.693 4.67e-08 ***
## Econ7.lag1          -1.256e-01  2.415e-02  -5.200 5.12e-07 ***
## Econ9.lag1          -5.877e-05  1.724e-05  -3.408 0.000799 ***
## Econ10.lag1         -1.822e-01  8.665e-02  -2.103 0.036806 *
## Econ11.lag1          1.695e-03  5.321e-04   3.185 0.001693 **
## Econ12.lag1         -6.526e-04  5.090e-04  -1.282 0.201368
## Econ13.lag1         -3.864e-04  1.159e-04  -3.333 0.001031 **
## Econ15.lag1          6.103e-01  1.268e-01   4.814 3.01e-06 ***
## Econ17.lag1         -5.790e-04  1.184e-04  -4.890 2.14e-06 ***
## Econ18.lag1          3.682e-05  1.722e-05   2.138 0.033817 *
## Econ19.lag1         -3.007e-06  1.232e-06  -2.440 0.015613 *
## Econ1.lag2          -2.088e-04  6.042e-05  -3.456 0.000675 ***
## Econ2.lag2           2.223e-01  9.331e-02   2.382 0.018183 *
## Econ3.lag2          -1.947e-01  4.156e-02  -4.684 5.34e-06 ***
## Econ5.lag2           4.634e-05  9.137e-06   5.072 9.32e-07 ***
## Econ6.lag2           6.159e-04  1.329e-04   4.636 6.60e-06 ***
## Econ7.lag2           1.048e-01  2.177e-02   4.814 3.01e-06 ***
```

```
## Econ8.lag2         -1.664e-02  3.197e-03  -5.207 4.97e-07 ***
## Econ11.lag2         1.116e-03  4.199e-04   2.658 0.008537 **
## Econ13.lag2        -2.868e-04  9.488e-05  -3.022 0.002854 **
## Econ15.lag2        -4.689e-01  1.134e-01  -4.136 5.31e-05 ***
## Econ16.lag2        -2.698e-01  6.716e-02  -4.017 8.48e-05 ***
## Econ18.lag2         9.468e-05  2.164e-05   4.375 2.00e-05 ***
## Econ19.lag2        -2.755e-06  1.185e-06  -2.325 0.021111 *
## Econ1.lag3         -2.824e-04  9.165e-05  -3.082 0.002365 **
## Econ2.lag3         -1.957e-01  5.479e-02  -3.572 0.000450 ***
## Econ3.lag3          3.788e-01  7.136e-02   5.308 3.08e-07 ***
## Econ4.lag3          1.151e-01  8.210e-02   1.402 0.162508
## Econ5.lag3         -2.753e-05  8.839e-06  -3.115 0.002122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2034 on 190 degrees of freedom
## Multiple R-squared:  0.9579, Adjusted R-squared:  0.9452
## F-statistic: 75.81 on 57 and 190 DF,  p-value: < 2.2e-16
```

As we see, the number of predictors has't decreased after executing the step(), this may mean that according to the AIC criterion, none of the predictors turned out to be insignificant enough to be excluded from the model.

Let's try using the regsubsets() function with the 57 predictors, that were defined in the previous steps, and with a maximum model size of 10 regressors.

```
best_subset <- regsubsets(y ~ START.YEAR + COMPLETION.YEAR + COMPLETION.QUARTER +
    PhysFin1 + PhysFin2 + PhysFin3 + PhysFin5 + PhysFin6 + PhysFin8 +
    Econ1 + Econ2 + Econ3 + Econ4 + Econ5 + Econ6 + Econ7 + Econ8 +
    Econ9 + Econ13 + Econ16 + Econ17 + Econ18 + Econ19 + Econ1.lag1 +
    Econ2.lag1 + Econ3.lag1 + Econ4.lag1 + Econ5.lag1 + Econ6.lag1 +
    Econ7.lag1 + Econ9.lag1 + Econ10.lag1 + Econ11.lag1 + Econ12.lag1 +
    Econ13.lag1 + Econ15.lag1 + Econ17.lag1 + Econ18.lag1 + Econ19.lag1 +
    Econ1.lag2 + Econ2.lag2 + Econ3.lag2 + Econ5.lag2 + Econ6.lag2 +
    Econ7.lag2 + Econ8.lag2 + Econ11.lag2 + Econ13.lag2 + Econ15.lag2 +
    Econ16.lag2 + Econ18.lag2 + Econ19.lag2 + Econ1.lag3 + Econ2.lag3 +
    Econ3.lag3 + Econ4.lag3 + Econ5.lag3, data = train_data, nvmax = 10, really.big = TRUE)
summary(best_subset)
```

```
## Subset selection object
## Call: regsubsets.formula(y ~ START.YEAR + COMPLETION.YEAR + COMPLETION.QUARTER +
##     PhysFin1 + PhysFin2 + PhysFin3 + PhysFin5 + PhysFin6 + PhysFin8 +
##     Econ1 + Econ2 + Econ3 + Econ4 + Econ5 + Econ6 + Econ7 + Econ8 +
##     Econ9 + Econ13 + Econ16 + Econ17 + Econ18 + Econ19 + Econ1.lag1 +
##     Econ2.lag1 + Econ3.lag1 + Econ4.lag1 + Econ5.lag1 + Econ6.lag1 +
##     Econ7.lag1 + Econ9.lag1 + Econ10.lag1 + Econ11.lag1 + Econ12.lag1 +
##     Econ13.lag1 + Econ15.lag1 + Econ17.lag1 + Econ18.lag1 + Econ19.lag1 +
##     Econ1.lag2 + Econ2.lag2 + Econ3.lag2 + Econ5.lag2 + Econ6.lag2 +
##     Econ7.lag2 + Econ8.lag2 + Econ11.lag2 + Econ13.lag2 + Econ15.lag2 +
##     Econ16.lag2 + Econ18.lag2 + Econ19.lag2 + Econ1.lag3 + Econ2.lag3 +
##     Econ3.lag3 + Econ4.lag3 + Econ5.lag3, data = train_data,
##     nvmax = 10, really.big = TRUE)
## 57 Variables  (and intercept)
##                    Forced in Forced out
## START.YEAR             FALSE      FALSE
## COMPLETION.YEAR        FALSE      FALSE
```

```
## COMPLETION.QUARTER     FALSE      FALSE
## PhysFin1               FALSE      FALSE
## PhysFin2               FALSE      FALSE
## PhysFin3               FALSE      FALSE
## PhysFin5               FALSE      FALSE
## PhysFin6               FALSE      FALSE
## PhysFin8               FALSE      FALSE
## Econ1                  FALSE      FALSE
## Econ2                  FALSE      FALSE
## Econ3                  FALSE      FALSE
## Econ4                  FALSE      FALSE
## Econ5                  FALSE      FALSE
## Econ6                  FALSE      FALSE
## Econ7                  FALSE      FALSE
## Econ8                  FALSE      FALSE
## Econ9                  FALSE      FALSE
## Econ13                 FALSE      FALSE
## Econ16                 FALSE      FALSE
## Econ17                 FALSE      FALSE
## Econ18                 FALSE      FALSE
## Econ19                 FALSE      FALSE
## Econ1.lag1             FALSE      FALSE
## Econ2.lag1             FALSE      FALSE
## Econ3.lag1             FALSE      FALSE
## Econ4.lag1             FALSE      FALSE
## Econ5.lag1             FALSE      FALSE
## Econ6.lag1             FALSE      FALSE
## Econ7.lag1             FALSE      FALSE
## Econ9.lag1             FALSE      FALSE
## Econ10.lag1            FALSE      FALSE
## Econ11.lag1            FALSE      FALSE
## Econ12.lag1            FALSE      FALSE
## Econ13.lag1            FALSE      FALSE
## Econ15.lag1            FALSE      FALSE
## Econ17.lag1            FALSE      FALSE
## Econ18.lag1            FALSE      FALSE
## Econ19.lag1            FALSE      FALSE
## Econ1.lag2             FALSE      FALSE
## Econ2.lag2             FALSE      FALSE
## Econ3.lag2             FALSE      FALSE
## Econ5.lag2             FALSE      FALSE
## Econ6.lag2             FALSE      FALSE
## Econ7.lag2             FALSE      FALSE
## Econ8.lag2             FALSE      FALSE
## Econ11.lag2            FALSE      FALSE
## Econ13.lag2            FALSE      FALSE
## Econ15.lag2            FALSE      FALSE
## Econ16.lag2            FALSE      FALSE
## Econ18.lag2            FALSE      FALSE
## Econ19.lag2            FALSE      FALSE
## Econ1.lag3             FALSE      FALSE
## Econ2.lag3             FALSE      FALSE
## Econ3.lag3             FALSE      FALSE
## Econ4.lag3             FALSE      FALSE
```

```
## Econ5.lag3             FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: exhaustive
##           START.YEAR COMPLETION.YEAR COMPLETION.QUARTER PhysFin1 PhysFin2
## 1  ( 1 )  " "        " "             " "                " "      " "
## 2  ( 1 )  " "        "*"             " "                "*"      " "
## 3  ( 1 )  " "        "*"             " "                "*"      " "
## 4  ( 1 )  "*"        " "             " "                "*"      " "
## 5  ( 1 )  "*"        " "             " "                "*"      " "
## 6  ( 1 )  "*"        " "             " "                "*"      " "
## 7  ( 1 )  "*"        " "             " "                "*"      " "
## 8  ( 1 )  "*"        " "             " "                "*"      " "
## 9  ( 1 )  "*"        " "             " "                "*"      " "
## 10  ( 1 ) "*"        "*"             "*"                "*"      " "
##           PhysFin3 PhysFin5 PhysFin6 PhysFin8 Econ1 Econ2 Econ3 Econ4 Econ5
## 1  ( 1 )  " "      " "      " "      "*"      " "   " "   " "   " "   " "
## 2  ( 1 )  " "      " "      " "      " "      " "   " "   " "   " "   " "
## 3  ( 1 )  " "      " "      " "      "*"      " "   " "   " "   " "   " "
## 4  ( 1 )  " "      " "      " "      "*"      " "   " "   " "   " "   " "
## 5  ( 1 )  " "      "*"      "*"      "*"      " "   " "   " "   " "   " "
## 6  ( 1 )  " "      "*"      "*"      "*"      " "   " "   " "   " "   " "
## 7  ( 1 )  " "      "*"      "*"      "*"      " "   " "   " "   " "   " "
## 8  ( 1 )  " "      "*"      "*"      "*"      " "   " "   "*"   " "   " "
## 9  ( 1 )  " "      "*"      "*"      "*"      " "   " "   "*"   " "   " "
## 10  ( 1 ) " "      "*"      "*"      "*"      " "   " "   "*"   " "   " "
##           Econ6 Econ7 Econ8 Econ9 Econ13 Econ16 Econ17 Econ18 Econ19 Econ1.lag1
## 1  ( 1 )  " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
## 2  ( 1 )  " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
## 3  ( 1 )  " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
## 4  ( 1 )  " "   " "   " "   " "   " "    "*"    " "    " "    " "    " "
## 5  ( 1 )  " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
## 6  ( 1 )  " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
## 7  ( 1 )  " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
## 8  ( 1 )  " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
## 9  ( 1 )  " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
## 10  ( 1 ) " "   " "   " "   " "   " "    " "    " "    " "    " "    " "
##           Econ2.lag1 Econ3.lag1 Econ4.lag1 Econ5.lag1 Econ6.lag1 Econ7.lag1
## 1  ( 1 )  " "        " "        " "        " "        " "        " "
## 2  ( 1 )  " "        " "        " "        " "        " "        " "
## 3  ( 1 )  " "        " "        " "        " "        " "        " "
## 4  ( 1 )  " "        " "        " "        " "        " "        " "
## 5  ( 1 )  " "        " "        " "        " "        " "        " "
## 6  ( 1 )  " "        " "        " "        " "        " "        " "
## 7  ( 1 )  " "        " "        " "        " "        " "        " "
## 8  ( 1 )  " "        " "        " "        " "        " "        "*"
## 9  ( 1 )  " "        " "        " "        " "        " "        "*"
## 10  ( 1 ) " "        " "        " "        " "        " "        "*"
##           Econ9.lag1 Econ10.lag1 Econ11.lag1 Econ12.lag1 Econ13.lag1
## 1  ( 1 )  " "        " "         " "         " "         " "
## 2  ( 1 )  " "        " "         " "         " "         " "
## 3  ( 1 )  " "        " "         " "         " "         " "
## 4  ( 1 )  " "        " "         " "         " "         " "
## 5  ( 1 )  " "        " "         " "         " "         " "
## 6  ( 1 )  " "        " "         " "         " "         " "
```

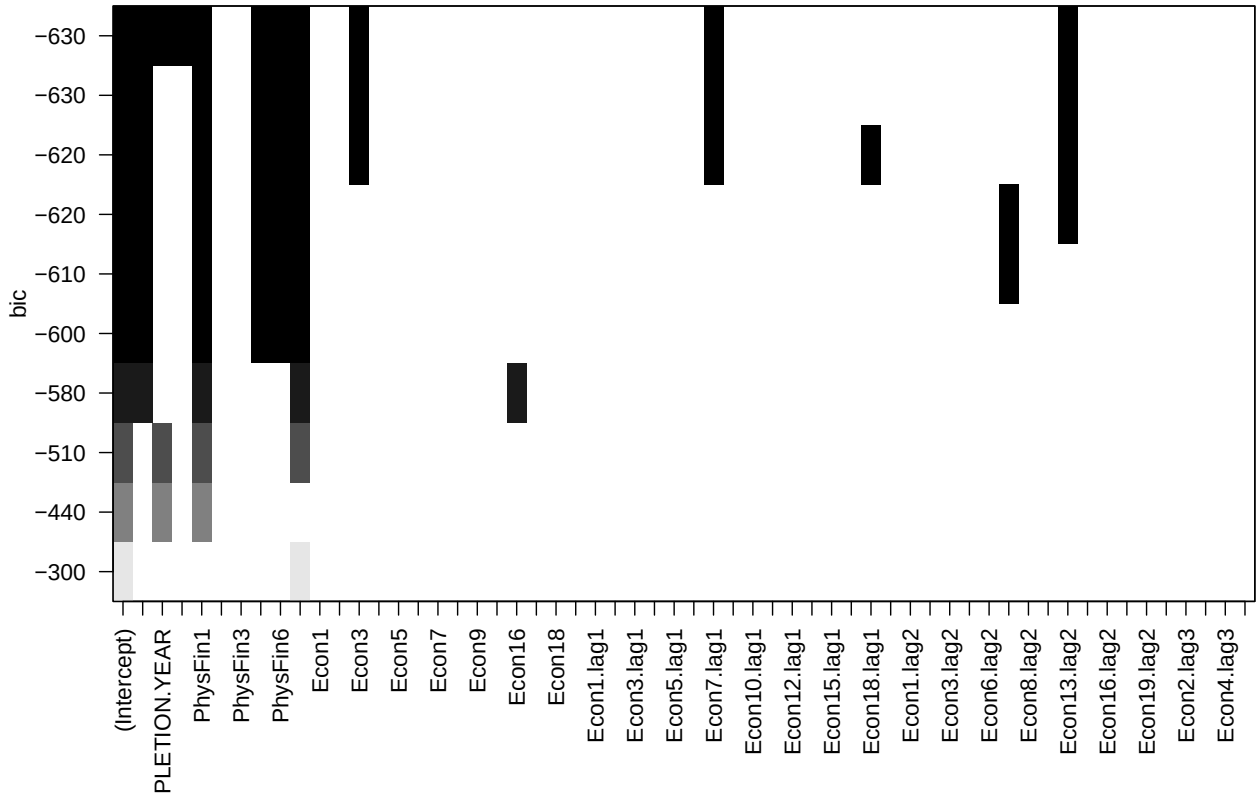
```
## 7  ( 1 )  " "         " "          " "          " "          " "
## 8  ( 1 )  " "         " "          " "          " "          " "
## 9  ( 1 )  " "         " "          " "          " "          " "
## 10  ( 1 ) " "         " "          " "          " "          " "
##           Econ15.lag1 Econ17.lag1 Econ18.lag1 Econ19.lag1 Econ1.lag2 Econ2.lag2
## 1  ( 1 )  " "          " "          " "          " "          " "         " "
## 2  ( 1 )  " "          " "          " "          " "          " "         " "
## 3  ( 1 )  " "          " "          " "          " "          " "         " "
## 4  ( 1 )  " "          " "          " "          " "          " "         " "
## 5  ( 1 )  " "          " "          " "          " "          " "         " "
## 6  ( 1 )  " "          " "          " "          " "          " "         " "
## 7  ( 1 )  " "          " "          " "          " "          " "         " "
## 8  ( 1 )  " "          " "          " "          " "          " "         " "
## 9  ( 1 )  " "          " "          "*"          " "          " "         " "
## 10  ( 1 ) " "          " "          " "          " "          " "         " "
##           Econ3.lag2 Econ5.lag2 Econ6.lag2 Econ7.lag2 Econ8.lag2 Econ11.lag2
## 1  ( 1 )  " "         " "         " "         " "         " "         " "
## 2  ( 1 )  " "         " "         " "         " "         " "         " "
## 3  ( 1 )  " "         " "         " "         " "         " "         " "
## 4  ( 1 )  " "         " "         " "         " "         " "         " "
## 5  ( 1 )  " "         " "         " "         " "         " "         " "
## 6  ( 1 )  " "         " "         " "         "*"         " "         " "
## 7  ( 1 )  " "         " "         " "         "*"         " "         " "
## 8  ( 1 )  " "         " "         " "         " "         " "         " "
## 9  ( 1 )  " "         " "         " "         " "         " "         " "
## 10  ( 1 ) " "         " "         " "         " "         " "         " "
##           Econ13.lag2 Econ15.lag2 Econ16.lag2 Econ18.lag2 Econ19.lag2
## 1  ( 1 )  " "          " "          " "          " "          " "
## 2  ( 1 )  " "          " "          " "          " "          " "
## 3  ( 1 )  " "          " "          " "          " "          " "
## 4  ( 1 )  " "          " "          " "          " "          " "
## 5  ( 1 )  " "          " "          " "          " "          " "
## 6  ( 1 )  " "          " "          " "          " "          " "
## 7  ( 1 )  "*"          " "          " "          " "          " "
## 8  ( 1 )  "*"          " "          " "          " "          " "
## 9  ( 1 )  "*"          " "          " "          " "          " "
## 10  ( 1 ) "*"          " "          " "          " "          " "
##           Econ1.lag3 Econ2.lag3 Econ3.lag3 Econ4.lag3 Econ5.lag3
## 1  ( 1 )  " "         " "         " "         " "         " "
## 2  ( 1 )  " "         " "         " "         " "         " "
## 3  ( 1 )  " "         " "         " "         " "         " "
## 4  ( 1 )  " "         " "         " "         " "         " "
## 5  ( 1 )  " "         " "         " "         " "         " "
## 6  ( 1 )  " "         " "         " "         " "         " "
## 7  ( 1 )  " "         " "         " "         " "         " "
## 8  ( 1 )  " "         " "         " "         " "         " "
## 9  ( 1 )  " "         " "         " "         " "         " "
## 10  ( 1 ) " "         " "         " "         " "         " "
```

We see that the function searched for the best subset for each size, starting from one predictor and up to 10 predictors. This means that she was looking for the best model with 1 predictor, with 2 predictors, and so on, up to 10 predictors.

As a result, we have the model with 10 predictors (last line): START.YEAR, COMPLETION.YEAR, COMPLETION.QUARTER, PhysFin1, PhysFin5, PhysFin6, PhysFin8, Econ3, Econ7.lag2, and Econ13.lag2.

## 2(b) Visualising the result of regsubsets() function

```
plot(best_subset, scale = "bic")
```




The most significant predictors are those with the most columns going up the Y axis. The graph shows that several predictors significantly lower the BIC value. This means that they have the greatest impact on improving the model.

In the graph, the names of the predictors are signed only through one. Upon closer examination, we see that the best model is with following predictors: START.YEAR, COMPLETION.YEAR, COMPLETION.QUARTER, PhysFin1, PhysFin5, PhysFin6, PhysFin8, Econ3, Econ7.lag2, and Econ13.lag2.

## 2(c) Applying lm() on the final best model and comparing the results with the previous ones(from (1b) and (1c)).

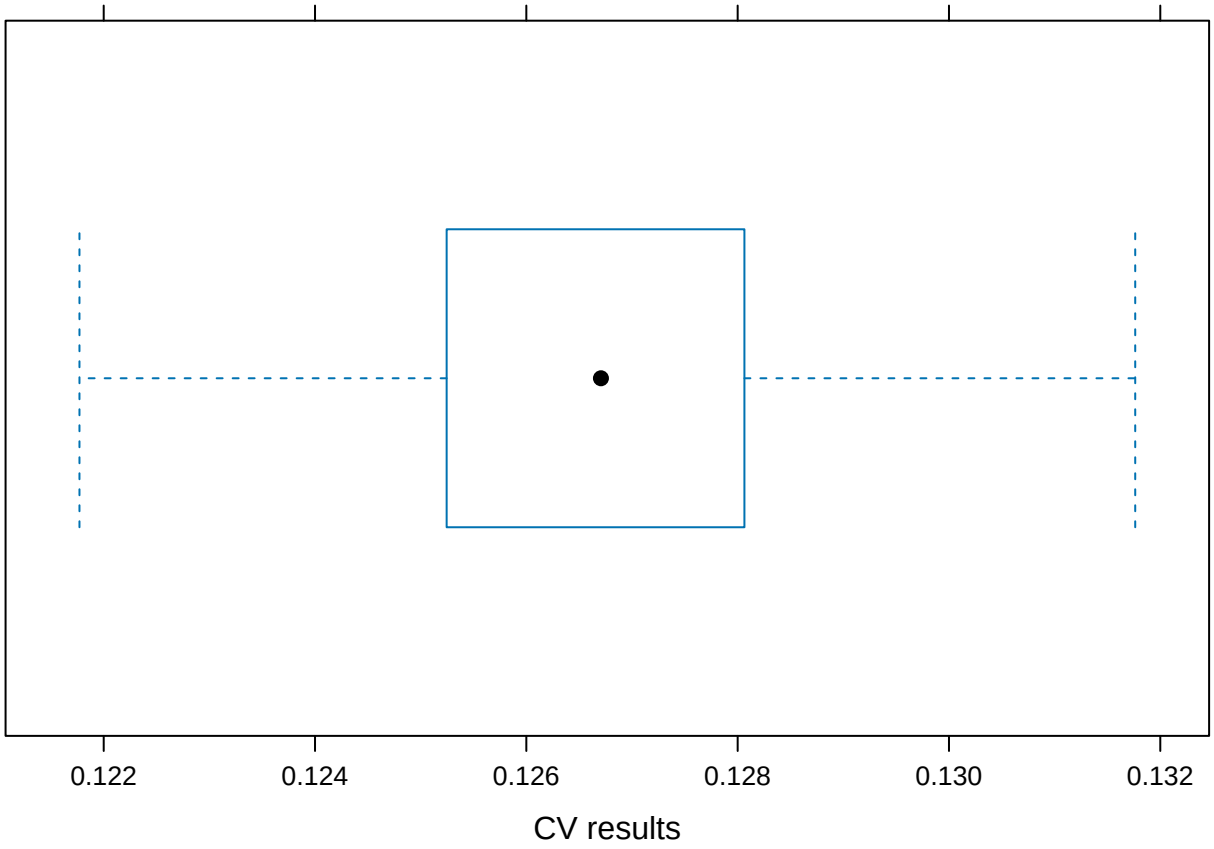
```
final_model <- lm(y ~ START.YEAR + COMPLETION.YEAR + COMPLETION.QUARTER +
                  PhysFin1 + PhysFin5 + PhysFin6 + PhysFin8 +
                  Econ3 + Econ7.lag2 + Econ13.lag2, data = train_data)

summary(final_model)
```

```
##
## Call:
## lm(formula = y ~ START.YEAR + COMPLETION.YEAR + COMPLETION.QUARTER +
##     PhysFin1 + PhysFin5 + PhysFin6 + PhysFin8 + Econ3 + Econ7.lag2 +
##     Econ13.lag2, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6727 -0.1187  0.0031  0.1330  0.6804
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -9.357e+00  8.279e-01 -11.301  < 2e-16 ***
## START.YEAR          1.015e-01  2.898e-02   3.503 0.000550 ***
## COMPLETION.YEAR     9.731e-02  2.756e-02   3.530 0.000499 ***
## COMPLETION.QUARTER  4.165e-02  1.537e-02   2.709 0.007242 **
## PhysFin1           -3.331e-02  3.464e-03  -9.618  < 2e-16 ***
## PhysFin5           -2.619e-03  4.331e-04  -6.047 5.71e-09 ***
## PhysFin6            5.049e-04  8.543e-05   5.910 1.18e-08 ***
## PhysFin8            4.470e-04  2.858e-05  15.641  < 2e-16 ***
## Econ3               4.393e-03  1.728e-03   2.541 0.011682 *
## Econ7.lag2         -4.749e-03  9.715e-04  -4.888 1.88e-06 ***
## Econ13.lag2        -3.505e-05  8.560e-06  -4.094 5.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2226 on 237 degrees of freedom
## Multiple R-squared:  0.9371, Adjusted R-squared:  0.9344
## F-statistic:   353 on 10 and 237 DF,  p-value: < 2.2e-16
```

The full model(from 1(a)) has a slightly better fit, as indicated by the smaller RSE and higher $R^2$ values. However, the difference with the reduced model is relatively small. Therefore, the reduced model is a good compromise between model fit quality and complexity, as reducing the number of predictors(from 73 to 10) didn't significantly affect the model's ability to explain the data.

```r
set.seed(123)
cv_result_final <- cvFit(final_model, data = train_data, y = train_data$y, cost = rtmspe,
                         K = 5, R = 100)
plot(cv_result_final)
```




Removing redundant features reduced variability and improved stability, resulting in more consistent cross-validation performance. The model now generalizes better.

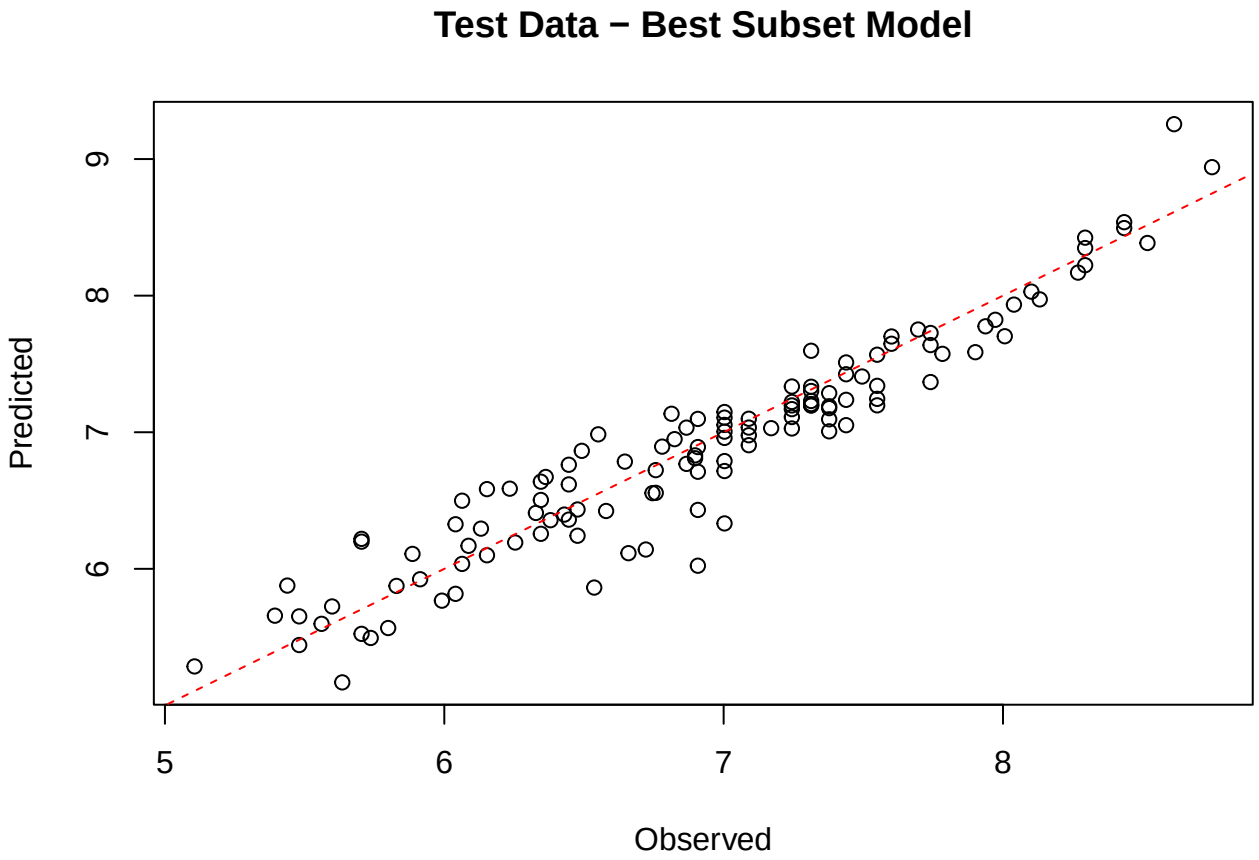
Notably, the median is centered within the box, indicating a symmetric distribution, and the narrow IQR suggests low variability. The whiskers show no significant outliers, and the absence of outliers points to more stable results compared to previous models (1(b),1(c)).

### 2(d) Predicting the response for the test data and visualising the result.

```
predicted_final_test <- predict(final_model, test_data)

rmse_final_test <- sqrt(mean((test_data$y - predicted_final_test)^2))
print(rmse_final_test)
```

```
## [1] 0.2520095
```

```
plot(test_data$y, predicted_final_test, xlab = 'Observed', ylab = 'Predicted',
     main = 'Test Data - Best Subset Model')
abline(0, 1, col = 'red', lty = 2)
```




The RMSE on the test data decreased significantly from 0.4190 for the full model to 0.2520 for the final model. This indicates that the final model, after feature selection, has a much lower prediction error on unseen data, suggesting better generalization and improved performance.

Some points on the plot deviate noticeably from the line, which indicates the presence of prediction errors. However, their number is less than in the case of the full model(1(d)), which is an improvement.