Ximan Liu
1935858
DSC 423
HW5

I have completed this work independently. The solutions given are entirely my own work.

1)
a)
We use K-fold cross validation to assess how well a model can be trained with certain data and then predict data it hasn't seen before. We can conduct k-fold cross validation with the 80/20 split by training the model k times on 80 percent of the data and testing on 20 percent. Each data point appears precisely once in the 20% test set. Cross-validation is used to assess models rather than construct them. We train the higher performing model after using cross-validation to find it. We don't utilize the actual model instances we trained during cross-validation in our final prediction model.

2)
a)
Initial first-order model (with all the variables):
PrizeMoney = -1174845.3 - 720.8 * AveDrivingDistance - 2457.9 * DrivingAccuracy + 10709.3 * GIR + 123072.1 * PuttingAverage + 11758.4 * BirdieConversion + 1074.7 * SandSaves + 4313.5 * Scrambling + 568.5 * BounceBack + 701.0 * PuttsPerRound + e

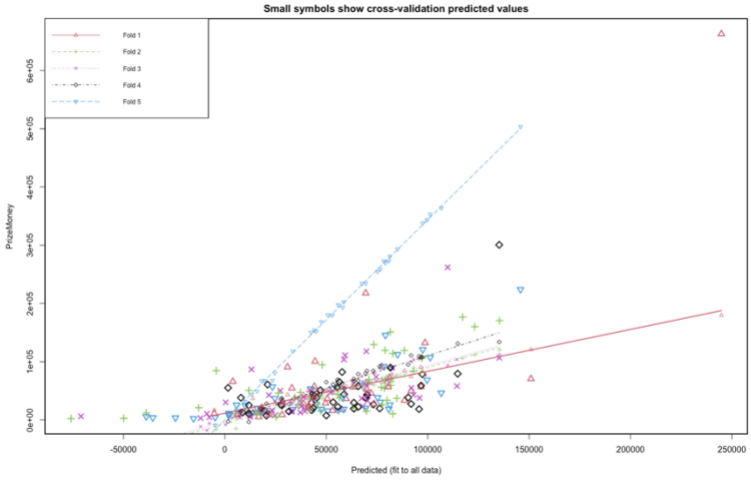
```
Call:
lm(formula = PrizeMoney ~ ., data = d)

Residuals:
   Min     1Q Median     3Q    Max
-80475 -26186  -6671  15209 417966

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1174845.3   591253.0  -1.987  0.04839 *
AveDrivingDistance     -720.8      766.0  -0.941  0.34795
DrivingAccuracy       -2457.9     1104.0  -2.226  0.02720 *
GIR                   10709.3     3761.8   2.847  0.00491 **
PuttingAverage       123072.1   579670.7   0.212  0.83209
BirdieConversion      11758.4     3801.9   3.093  0.00229 **
SandSaves              1074.7      759.4   1.415  0.15868
Scrambling             4313.5     2477.8   1.741  0.08336 .
BounceBack              568.5     1585.2   0.359  0.72028
PuttsPerRound           701.0    37306.6   0.019  0.98503
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50270 on 186 degrees of freedom
Multiple R-squared:  0.4097,    Adjusted R-squared:  0.3811
F-statistic: 14.34 on 9 and 186 DF,  p-value: < 2.2e-16
```

 # 5 cross-validation (initial first-order model with all the variables):

out <- cv.lm(data = d, form.lm = formula(PrizeMoney ~ .), plotit = "Observed", m=5)

```
Analysis of Variance Table

Response: PrizeMoney
                    Df   Sum Sq  Mean Sq F value  Pr(>F)
AveDrivingDistance   1 2.01e+10 2.01e+10    7.97 0.00529 **
DrivingAccuracy      1 1.72e+10 1.72e+10    6.80 0.00983 **
GIR                  1 1.18e+11 1.18e+11   46.83 1.1e-10 ***
PuttingAverage       1 9.51e+10 9.51e+10   37.64 5.0e-09 ***
BirdieConversion     1 2.86e+10 2.86e+10   11.32 0.00093 ***
SandSaves            1 2.53e+10 2.53e+10   10.01 0.00182 **
Scrambling           1 2.12e+10 2.12e+10    8.38 0.00425 **
BounceBack           1 3.31e+08 3.31e+08    0.13 0.71802
PuttsPerRound        1 8.92e+05 8.92e+05    0.00 0.98503
Residuals          186 4.70e+11 2.53e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
fold 5
Observations in test set: 39
                3       6      16      22    23      24      27
Predicted   -24357  101255   81342   97559 10061   45069   44146
cvpred      -79613  352822  280641  338058 31512  151986  154291
PrizeMoney    3635  107294   26129  120927 24814   27224   20322
CV residual  83248 -245528 -254512 -217131 -6698 -124762 -133969
               29      31      32     40      45     46      67
Predicted   67591   50721   85132  23442  106678  19816   76572
cvpred     234287  180405  293376  67770  363670  67441  258476
PrizeMoney  60073   15668  112443  56873   46377  16630   30656
CV residual -174214 -164737 -180933 -10897 -317293 -50811 -227820
              76      82     99    101    103     112     114
Predicted    5735   55846  42468 -15440  33505   58588   47519
cvpred      17059  197488 150092 -53877 117813  202308  168439
PrizeMoney  25804   16927  53530   2426  18085   18494   18721
CV residual  8745 -180561 -96562  56303 -99728 -183814 -149718
```

```
                118     119     126   127   131    139     142
Predicted     69393  -38757   80849 -4762  1870  24663  145751
cvpred       234311 -134111  272964 -8955 13164  80436  503399
PrizeMoney    20188    5777   18838  4444  8272  37100  224027
CV residual -214123  139888 -254126 13399 -4892 -43336 -279372
                143     145     154     161     164   170    181
Predicted     79145   52949  -35509   78381   75298  1985  15880
cvpred       269610  179719 -119819  273049  254634  4276  49135
PrizeMoney   145414   53634    3816   91808   38471 11421  20064
CV residual -124196 -126085  123635 -181241 -216163  7145 -29071
                183    187     191     195
Predicted     13250  18116   99638   58089
cvpred        43766  67048  343902  193205
PrizeMoney    11309  14098   68613   38043
CV residual -32457 -52950 -275289 -155162

Sum of squares = 1.01e+12    Mean square = 2.59e+10    n = 39

Overall (Sum over all 39 folds)
      ms
7.87e+09
```



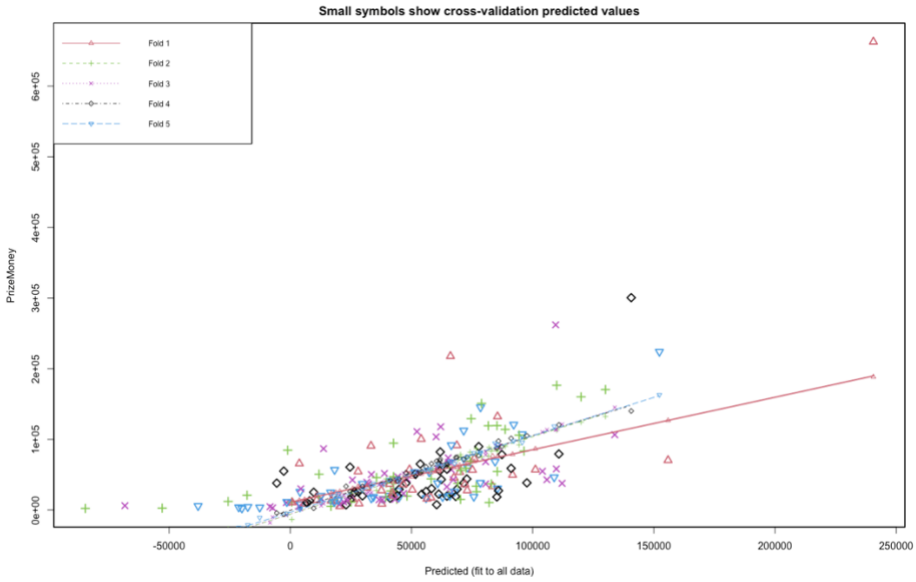
Final first-order model (after remove variables one-by-one):
PrizeMoney = -1094996.9 - 1964.1 * DrivingAccuracy + 9742.9 * GIR + 10670.5 * BirdieConversion + 5670.4 * Scrambling + e

```
Call:
lm(formula = PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion +
    Scrambling, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-85429 -27959  -7833  15674 422173

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1094996.9   109585.4  -9.992  < 2e-16 ***
DrivingAccuracy     -1964.1      815.7  -2.408    0.017 *
GIR                  9742.9     1465.9   6.646 3.06e-10 ***
BirdieConversion    10670.5     1703.7   6.263 2.44e-09 ***
Scrambling           5670.4     1239.4   4.575 8.56e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50080 on 191 degrees of freedom
Multiple R-squared:  0.3984,    Adjusted R-squared:  0.3858
F-statistic: 31.62 on 4 and 191 DF,  p-value: < 2.2e-16
```

 # 5 cross-validation (final first-order model after remove variables one-by-one):
out <- cv.lm(data = d, form.lm = formula(PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + Scrambling), plotit = "Observed", m=5)

```
Analysis of Variance Table

Response: PrizeMoney
                  Df   Sum Sq  Mean Sq F value  Pr(>F)
DrivingAccuracy    1 4.85e+08 4.85e+08    0.19    0.66
GIR                1 1.54e+11 1.54e+11   61.43 3.1e-13 ***
BirdieConversion   1 1.10e+11 1.10e+11   43.92 3.4e-10 ***
Scrambling         1 5.25e+10 5.25e+10   20.93 8.6e-06 ***
Residuals        191 4.79e+11 2.51e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
fold 5
Observations in test set: 39
                3     6     16     22    23     24     27     29
Predicted   -12615 95669  67755  92159 16635  44308  40707  73159
cvpred      -10808 95726  65613  97721 10928  44698  42947  75646
PrizeMoney    3635 107294  26129 120927 24814  27224  20322  60073
CV residual  14443 11568 -39484  23206 13886 -17474 -22625 -15573
               31    32    40     45     46     67    76     82
Predicted    38506 71536 18275 108853  33406  85467  4012  56013
cvpred       36402 69503  9888 118549  33953  92210   712  57418
PrizeMoney   15668 112443 56873  46377  16630  30656 25804  16927
CV residual -20734 42940 46985 -72172 -17323 -61554 25092 -40491
              99    101    103    112    114    118    119    126
Predicted   51334 -19957  44849  62841  34304  65781 -38049  75724
cvpred      55129 -27603  39742  62088  37224  69136 -44782  80774
PrizeMoney  53530   2426  18085  18494  18721  20188   5777  18838
CV residual -1599  30029 -21657 -43594 -18503 -48948  50559 -61936
              127  131   139    142    143   145    154   161
Predicted   -17537 4170 29749 152277  78470 57312 -21267 66430
cvpred      -21112 2516 28218 162993  80257 57393 -24769 68989
PrizeMoney    4444 8272 37100 224027 145414 53634   3816 91808
CV residual  25556 5756  8882  61034  65157 -3759  28585 22819
```

```
              164   170   181   183   187    191    195
Predicted   78409 -1424 12640 18148 19818  84465  60791
cvpred      81993 -4872 11055 20070 23112  94356  63937
PrizeMoney  38471 11421 20064 11309 14098  68613  38043
CV residual -43522 16293  9009 -8761 -9014 -25743 -25894

Sum of squares = 4.6e+10    Mean square = 1.18e+09    n = 39

Overall (Sum over all 39 folds)
      ms
2.82e+09
```



b)

Final second-order model (after remove variables one-by-one):

PrizeMoney = (1.25e+07) + (-2.03e+05) * GIR + (-2.48e+05) * BirdieConversion + (-1.16e+05) * SandSaves + (-4.19e+01) * ADD2 + (2.67e+02) * DA2 + (8.70e+02) * G2 + (7.46e+02) * BB2 + (3.31e+02) * ADDBC + (2.94e+02) * ADDSS + (-5.52e+02) * DAG + (3.47e+04) * GPA + (2.50e+03) * GBC + (-2.26e+04) * PASB + (-4.37e+04) * PABB + (5.60e+02) * SSSB + (8.54e+02) * SBBB + e

P.S.:

```
d$ADD2 <- d$AveDrivingDistance^2
d$DA2 <- d$DrivingAccuracy^2
d$G2 <- d$GIR^2
d$BB2 <- d$BounceBack^2
d$ADDBC <- d$AveDrivingDistance * d$BirdieConversion
d$ADDSS <- d$AveDrivingDistance * d$SandSaves
d$DAG <- d$DrivingAccuracy * d$GIR
d$GPA <- d$GIR * d$PuttingAverage
d$GBC <- d$GIR * d$BirdieConversion
```

d$PASB <- d$PuttingAverage * d$Scrambling
d$PABB <- d$PuttingAverage * d$BounceBack
d$SSSB <- d$SandSaves * d$Scrambling
d$SBBB <- d$Scrambling * d$BounceBack

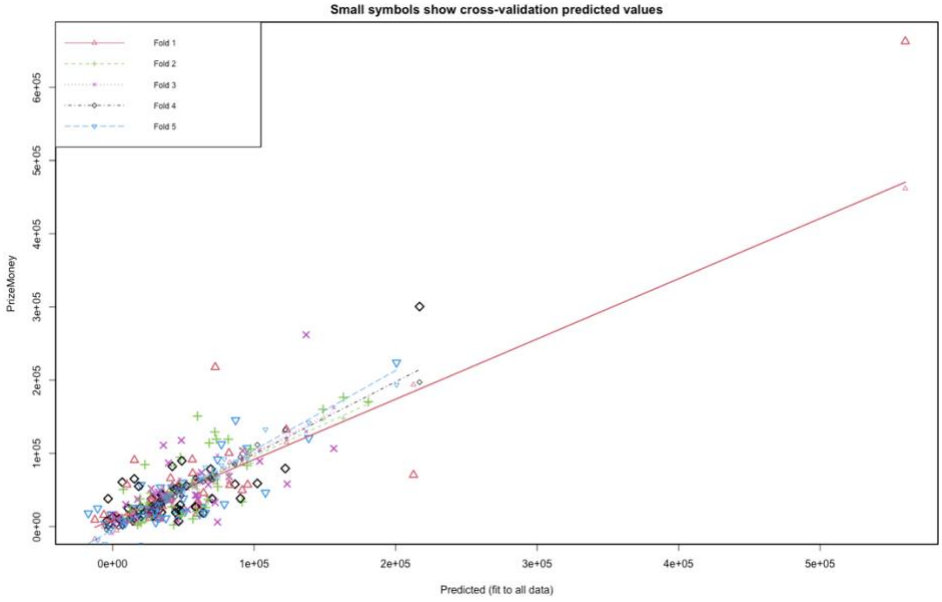
```
Call:
lm(formula = PrizeMoney ~ GIR + BirdieConversion + SandSaves +
    ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DAG + GPA + GBC +
    PASB + PABB + SSSB + SBBB, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-142234  -19973    -990   12435  145195

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.25e+07   1.21e+06   10.29  < 2e-16 ***
GIR             -2.03e+05   2.97e+04   -6.83  1.3e-10 ***
BirdieConversion -2.48e+05   4.58e+04   -5.41  2.0e-07 ***
SandSaves       -1.16e+05   2.31e+04   -5.03  1.2e-06 ***
ADD2            -4.19e+01   9.14e+00   -4.58  8.7e-06 ***
DA2              2.67e+02   8.54e+01    3.13  0.00202 **
G2               8.70e+02   2.50e+02    3.48  0.00062 ***
BB2              7.46e+02   2.59e+02    2.88  0.00442 **
ADDBC            3.31e+02   1.65e+02    2.00  0.04709 *
ADDSS            2.94e+02   6.29e+01    4.68  5.6e-06 ***
DAG             -5.52e+02   1.69e+02   -3.26  0.00134 **
GPA              3.47e+04   7.52e+03    4.61  7.7e-06 ***
GBC              2.50e+03   4.15e+02    6.02  9.5e-09 ***
PASB            -2.26e+04   4.75e+03   -4.75  4.2e-06 ***
PABB            -4.37e+04   1.12e+04   -3.92  0.00013 ***
SSSB             5.60e+02   1.34e+02    4.19  4.3e-05 ***
SBBB             8.54e+02   3.05e+02    2.80  0.00567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35800 on 179 degrees of freedom
Multiple R-squared:  0.712,     Adjusted R-squared:  0.686
F-statistic: 27.7 on 16 and 179 DF,  p-value: <2e-16
```

 # 5 cross-validation (final second-order model after remove variables one-by-one):
out <- cv.lm(data = d, form.lm = formula(PrizeMoney ~ GIR + BirdieConversion + SandSaves + ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DAG + GPA + GBC + PASB + PABB + SSSB + SBBB), plotit = "Observed", m=5)

```
Analysis of Variance Table

Response: PrizeMoney
                  Df   Sum Sq  Mean Sq F value  Pr(>F)
GIR                1 1.34e+11 1.34e+11  104.64 < 2e-16 ***
BirdieConversion   1 1.29e+11 1.29e+11  100.77 < 2e-16 ***
SandSaves          1 3.32e+10 3.32e+10   25.93 8.9e-07 ***
ADD2               1 1.24e+02 1.24e+02    0.00 0.99975
DA2                1 7.13e+09 7.13e+09    5.57 0.01934 *
G2                 1 7.84e+10 7.84e+10   61.22 4.3e-13 ***
BB2                1 3.46e+09 3.46e+09    2.70 0.10198
ADDBC              1 3.40e+10 3.40e+10   26.52 6.8e-07 ***
ADDSS              1 6.91e+09 6.91e+09    5.39 0.02133 *
DAG                1 1.77e+10 1.77e+10   13.85 0.00026 ***
GPA                1 8.06e+07 8.06e+07    0.06 0.80215
GBC                1 6.14e+10 6.14e+10   47.92 7.6e-11 ***
PASB               1 1.61e+10 1.61e+10   12.61 0.00049 ***
PABB               1 9.96e+09 9.96e+09    7.78 0.00586 **
SSSB               1 2.56e+10 2.56e+10   20.03 1.4e-05 ***
SBBB               1 1.00e+10 1.00e+10    7.84 0.00567 **
Residuals        179 2.29e+11 1.28e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
fold 5
Observations in test set: 39
                3      6     16     22     23     24    27    29
Predicted   -2314  95059  63431 138588 -10547  62842 24542 49806
cvpred      -7244  92875  69197 141531 -18360  63349 25095 46428
PrizeMoney   3635 107294  26129 120927  24814  27224 20322 60073
CV residual 10879  14419 -43068 -20604  43174 -36125 -4773 13645
               31     32     40     45     46     67    76     82
Predicted    45593  76894  20108 108018 -1000  79003 14177  61490
cvpred       51181  83765 -24568 132599 -3795  92282  9861  65115
PrizeMoney   15668 112443  56873  46377 16630  30656 25804  16927
CV residual -35513  28678  81441 -86222 20425 -61626 15943 -48188
              99   101    103    112   114    118    119    126
Predicted   33877  7153 -17156  47196 18577  51062  30520  65515
cvpred      30558  4676 -39960  49111 27468  60386  31339  80999
PrizeMoney  53530  2426  18085  18494 18721  20188   5777  18838
CV residual 22972 -2250  58045 -30617 -8747 -40198 -25562 -62161
             127  131   139    142    143   145    154   161
Predicted    8915 5914 37205 200661  86930 41904  -5502 74202
cvpred       9844 8949 38126 193753  88540 43404 -23614 66838
PrizeMoney   4444 8272 37100 224027 145414 53634   3816 91808
CV residual -5400 -677 -1026  30274  56874 10230  27430 24970
```

```
              164    170   181    183    187   191   195
Predicted   50027  31937 30003  37691  20899 71526 38401
cvpred      53769  29315 18157  42288  26574 77853 41197
PrizeMoney  38471  11421 20064  11309  14098 68613 38043
CV residual -15298 -17894  1907 -30979 -12476 -9240 -3154

Sum of squares = 4.74e+10    Mean square = 1.22e+09    n = 39

Overall (Sum over all 39 folds)
      ms
1.62e+09
```



c)

The final second-order model (backward selection) reaches different result from the above one. It has one more variable called "DrivingAccuracy".

Final second-order model (backward selection):
PrizeMoney = (1.18e+07) + (2.02e+04) * DrivingAccuracy + (-2.10e+05) * GIR + (-2.35e+05) * BirdieConversion + (-1.12e+05) * SandSaves + (-4.09e+01) * ADD2 + (2.65e+02) * DA2 + (1.10e+03) * G2 + (7.51e+02) * BB2 + (3.33e+02) * ADDBC + (2.83e+02) * ADDSS + (-8.54e+02) * DAG + (3.59e+04) * GPA + (2.28e+03) * GBC + (-2.33e+04) * PASB + (-4.72e+04) * PABB + (5.42e+02) * SSSB + (9.53e+02) * SBBB + e

P.S.:

```
d$ADD2 <- d$AveDrivingDistance^2
d$DA2 <- d$DrivingAccuracy^2
d$G2 <- d$GIR^2
d$BB2 <- d$BounceBack^2
d$ADDBC <- d$AveDrivingDistance * d$BirdieConversion
d$ADDSS <- d$AveDrivingDistance * d$SandSaves
d$DAG <- d$DrivingAccuracy * d$GIR
```

d$GPA <- d$GIR * d$PuttingAverage
d$GBC <- d$GIR * d$BirdieConversion
d$PASB <- d$PuttingAverage * d$Scrambling
d$PABB <- d$PuttingAverage * d$BounceBack
d$SSSB <- d$SandSaves * d$Scrambling
d$SBBB <- d$Scrambling * d$BounceBack

```
> # Display results
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
PrizeMoney ~ AveDrivingDistance + DrivingAccuracy + GIR + PuttingAverag
e +
    BirdieConversion + SandSaves + Scrambling + BounceBack +
    PuttsPerRound + ADD2 + DA2 + G2 + PA2 + BC2 + SS2 + SB2 +
    BB2 + PPR2 + ADDG + ADDPA + ADDBC + ADDSS + ADDBB + DAG +
    DAPA + DASB + GPA + GBC + GSB + GPPR + PASB + PASS + PABB +
    SSSB + SBBB

Final Model:
PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + SandSaves +
    ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DAG + GPA + GBC +
    PASB + PABB + SSSB + SBBB

                   Step Df Deviance Resid. Df Resid. Dev  AIC
1                                         160   2.21e+11 4157
2                 - GPPR  1 1.40e+07       161   2.21e+11 4155
3                 - DAPA  1 7.57e+07       162   2.21e+11 4153
4                - ADDPA  1 5.82e+07       163   2.21e+11 4151
5                  - PA2  1 9.86e+07       164   2.21e+11 4149
6  - AveDrivingDistance  1 2.14e+08       165   2.21e+11 4147
7      - PuttingAverage  1 1.93e+08       166   2.21e+11 4145
8          - Scrambling  1 2.29e+08       167   2.21e+11 4144
9          - BounceBack  1 3.32e+08       168   2.22e+11 4142
10               - ADDBB  1 2.69e+08       169   2.22e+11 4140
11                 - SB2  1 2.80e+08       170   2.22e+11 4138
12                - DASB  1 2.71e+08       171   2.23e+11 4137
13                 - SS2  1 3.02e+08       172   2.23e+11 4135
14                - ADDG  1 3.65e+08       173   2.23e+11 4133
15                 - GSB  1 7.53e+08       174   2.24e+11 4132
16                 - BC2  1 7.04e+08       175   2.25e+11 4131
17                - PPR2  1 3.37e+08       176   2.25e+11 4129
18      - PuttsPerRound  1 6.38e+08       177   2.26e+11 4127
19                - PASS  1 1.04e+09       178   2.27e+11 4126
```

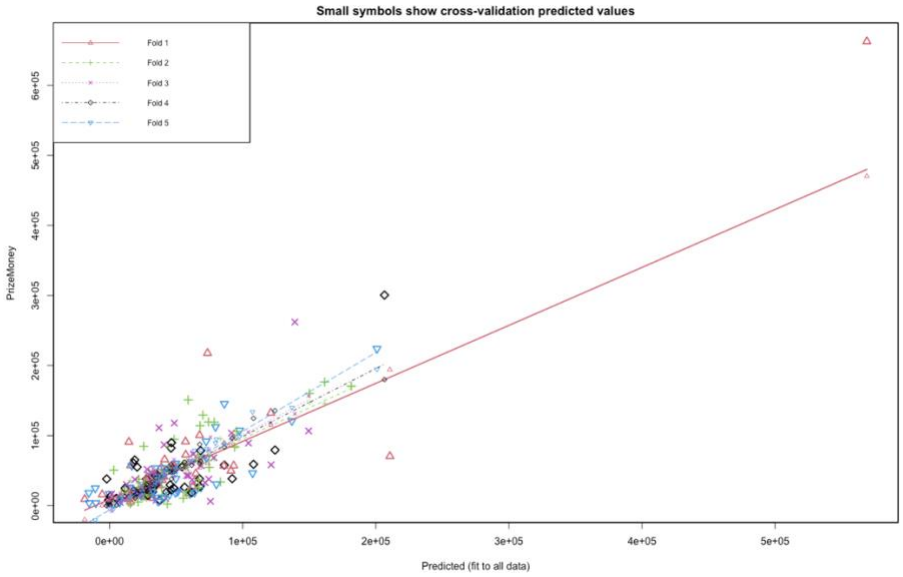
```
Call:
lm(formula = PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion +
    SandSaves + ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DAG +
    GPA + GBC + PASB + PABB + SSSB + SBBB, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-140142  -20736    -444   14039  144209

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.18e+07   1.30e+06    9.06  2.2e-16 ***
DrivingAccuracy  2.02e+04   1.44e+04    1.40   0.1633
GIR             -2.10e+05   3.00e+04   -6.99  5.4e-11 ***
BirdieConversion -2.35e+05   4.67e+04   -5.02  1.3e-06 ***
SandSaves       -1.12e+05   2.33e+04   -4.81  3.2e-06 ***
ADD2            -4.09e+01   9.15e+00   -4.47  1.4e-05 ***
DA2              2.65e+02   8.52e+01    3.11   0.0022 **
G2               1.10e+03   2.99e+02    3.68   0.0003 ***
BB2              7.51e+02   2.58e+02    2.91   0.0041 **
ADDBC            3.33e+02   1.65e+02    2.02   0.0454 *
ADDSS            2.83e+02   6.33e+01    4.48  1.4e-05 ***
DAG             -8.54e+02   2.74e+02   -3.12   0.0021 **
GPA              3.59e+04   7.55e+03    4.75  4.1e-06 ***
GBC              2.28e+03   4.44e+02    5.13  7.4e-07 ***
PASB            -2.33e+04   4.77e+03   -4.89  2.3e-06 ***
PABB            -4.72e+04   1.14e+04   -4.14  5.4e-05 ***
SSSB             5.42e+02   1.34e+02    4.05  7.6e-05 ***
SBBB             9.53e+02   3.12e+02    3.05   0.0026 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35700 on 178 degrees of freedom
Multiple R-squared:  0.715,     Adjusted R-squared:  0.688
F-statistic: 26.3 on 17 and 178 DF,  p-value: <2e-16
```

 # 5 cross-validation(final second-order model - backward selection):
out <- cv.lm(data = d, form.lm = formula(PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + SandSaves + ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DAG + GPA + GBC + PASB + PABB + SSSB + SBBB), plotit = "Observed", m=5)

```
Analysis of Variance Table

Response: PrizeMoney
                 Df   Sum Sq  Mean Sq F value  Pr(>F)
DrivingAccuracy   1 4.85e+08 4.85e+08    0.38  0.5380
GIR               1 1.54e+11 1.54e+11  120.97 < 2e-16 ***
BirdieConversion  1 1.10e+11 1.10e+11   86.49 < 2e-16 ***
SandSaves         1 3.56e+10 3.56e+10   27.92 3.7e-07 ***
ADD2              1 4.02e+09 4.02e+09    3.16  0.0772 .
DA2               1 2.88e+09 2.88e+09    2.26  0.1343
G2                1 7.46e+10 7.46e+10   58.56 1.2e-12 ***
BB2               1 3.52e+09 3.52e+09    2.77  0.0980 .
ADDBC             1 3.40e+10 3.40e+10   26.66 6.5e-07 ***
ADDSS             1 7.24e+09 7.24e+09    5.68  0.0182 *
DAG               1 3.27e+10 3.27e+10   25.69 1.0e-06 ***
GPA               1 2.91e+08 2.91e+08    0.23  0.6333
GBC               1 4.82e+10 4.82e+10   37.85 4.9e-09 ***
PASB              1 1.48e+10 1.48e+10   11.63  0.0008 ***
PABB              1 1.02e+10 1.02e+10    8.03  0.0051 **
SSSB              1 2.49e+10 2.49e+10   19.57 1.7e-05 ***
SBBB              1 1.19e+10 1.19e+10    9.32  0.0026 **
Residuals       178 2.27e+11 1.27e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
fold 5
Observations in test set: 39
                3      6     16     22     23     24    27    29
Predicted   -10358  97593  68650 136949 -10807  62869 23992 49449
cvpred      -25054  98419  79348 139822 -20652  63710 24268 45118
PrizeMoney    3635 107294  26129 120927  24814  27224 20322 60073
CV residual  28689   8875 -53219 -18895  45466 -36486 -3946 14955
               31     32     40     45     46     67    76     82
Predicted    44913  79533  16677 107423  -194  80206 15022  60325
cvpred       50843  89374 -31239 133823 -3700  94844 11989  64061
PrizeMoney   15668 112443  56873  46377 16630  30656 25804  16927
CV residual -35175  23069  88112 -87446 20330 -64188 13815 -47134
              99   101    103    112   114    118    119    126
Predicted   34101  2912 -15654  49495 17068  51150  36517  63766
cvpred      30079 -4627 -38980  52371 26151  61522  42748  78761
PrizeMoney  53530  2426  18085  18494 18721  20188   5777  18838
CV residual 23451  7053  57065 -33877 -7430 -41334 -36971 -59923
             127   131   139    142    143   145    154   161
Predicted    7491  8534 36045 200877  86174 41199 -15390 72508
cvpred       8811 12089 37974 194701  87473 43216 -46525 64620
PrizeMoney   4444  8272 37100 224027 145414 53634   3816 91808
CV residual -4367 -3817  -874  29326  57941 10418  50341 27188
```

```
              164    170   181    183    187    191   195
Predicted    49777  34316 32484  40845  23792  72998 38695
cvpred       53499  33743 20526  48235  31697  81231 41803
PrizeMoney   38471  11421 20064  11309  14098  68613 38043
CV residual -15028 -22322  -462 -36926 -17599 -12618 -3760

Sum of squares = 5.38e+10    Mean square = 1.38e+09    n = 39

Overall (Sum over all 39 folds)
      ms
1.64e+09
```



d)

Final second-order model (after remove variables one-by-one):

PrizeMoney = (1.25e+07) + (-2.03e+05) * GIR + (-2.48e+05) * BirdieConversion + (-1.16e+05) * SandSaves + (-4.19e+01) * ADD2 + (2.67e+02) * DA2 + (8.70e+02) * G2 + (7.46e+02) * BB2 + (3.31e+02) * ADDBC + (2.94e+02) * ADDSS + (-5.52e+02) * DAG + (3.47e+04) * GPA + (2.50e+03) * GBC + (-2.26e+04) * PASB + (-4.37e+04) * PABB + (5.60e+02) * SSSB + (8.54e+02) * SBBB + e

Final second-order model (backward selection):

PrizeMoney = (1.18e+07) + (2.02e+04) * DrivingAccuracy + (-2.10e+05) * GIR + (-2.35e+05) * BirdieConversion + (-1.12e+05) * SandSaves + (-4.09e+01) * ADD2 + (2.65e+02) * DA2 + (1.10e+03) * G2 + (7.51e+02) * BB2 + (3.33e+02) * ADDBC + (2.83e+02) * ADDSS + (-8.54e+02) * DAG + (3.59e+04) * GPA + (2.28e+03) * GBC + (-2.33e+04) * PASB + (-4.72e+04) * PABB + (5.42e+02) * SSSB + (9.53e+02) * SBBB + e

These two models are almost the same except the "removing variables one-by-one" does not have the variable "DrivingAccuracy" like the second model, as known as the backward selection one. The model without "DrivingAccuracy" is better.

There is no doubt that the stepwise backward selection way to identify final model is so much quicker than the first one. However, the "DrivingAccuracy" is best not to add this variable to the final model because the p-value of it is larger than a=0.05. Using scatter plots to delete irrelevant variables one by one may be very slow, but it is more accurate in this problem. The variables determined by stepwise backward selection can explain the variation of the dependent variable appropriately, but the negative effect of stepwise selection is to bias the model, which so-called overfitting problem.