DSC 423: Data Analysis and Regression

Assignment 5:

Name: Adarsh Shankar

Student ID: 2117611

Honor Statement: "I have completed this work independently. The solutions given are entirely my own work."

Question 1: Short Essay. The purpose of k-fold cross validation is often misunderstood. a. (10 points) How do you use cross validation to select a final (or production) model? Note: it is not the "best" of the k models you have built using cross validation.

Answer: K-fold cross validation is a method for evaluating how well a model can be trained on a set of data and then used to predict new data. We can perform k-fold cross validation using the 80/20 split by training the model k times on 80% of the data and testing on the remaining 20%. The 20% test set contained Each piece of information only appears once. Instead of creating them from scratch, the technique known as cross-validation is utilised to evaluate them. Cross-validation is used to determine which model performs better before we educate it. We don't employ the model instances we trained during cross-validation in our final prediction model.

Question 2: PGA. The pgatour2006.csv dataset contains data for 196 players. The variables in the dataset are:

- Player's name
- PrizeMoney = average prize money per tournament
- DrivingAccuracy = percent of times a player is able to hit the fairway with his tee shot
- GIR = percent of time a player was able to hit the green within two or less than par (Greens in Regulation)
- BirdieConversion = percentage of times a player makes a birdie or better after hitting the green in regulation
- PuttingAverage = putting performance on those holes where the green was hit in regulation.
- PuttsPerRound= average number of putts per round (shots played on the green)
- Etc.

a. (10 points) Build a complete first-order model. Evaluate the model using 5-fold cross validation. If necessary, remove a non-significant variable and repeat until you have your final first-order model. Present the model.

Solution:

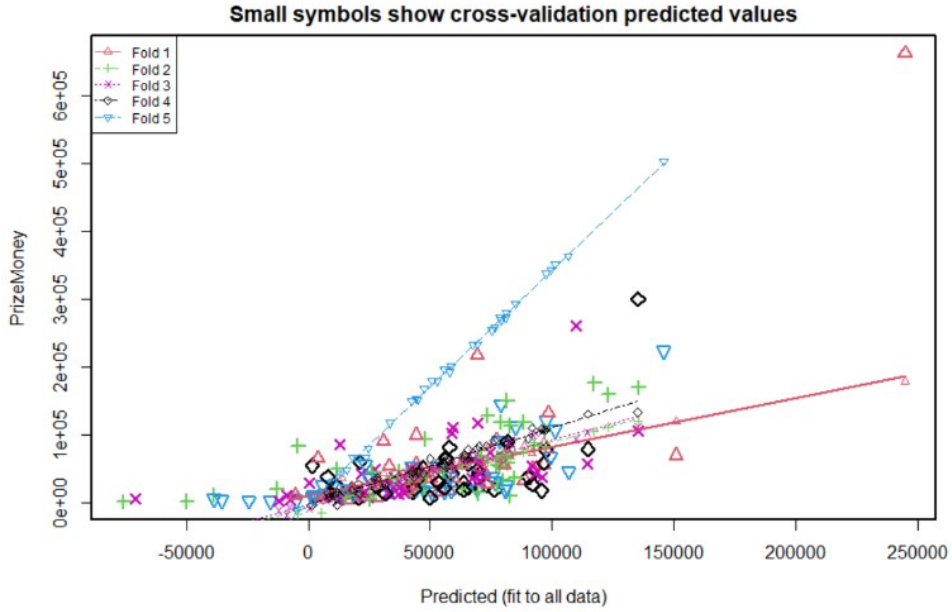
Initialize first model with all the variables

```
> model0 <- lm(PrizeMoney ~ ., data = tour)
> summary(model0)

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

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

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1174845     591253   -1.99   0.0484 *
AveDrivingDistance     -721        766   -0.94   0.3480
DrivingAccuracy       -2458       1104   -2.23   0.0272 *
GIR                   10709       3762    2.85   0.0049 **
PuttingAverage       123072     579671    0.21   0.8321
BirdieConversion      11758       3802    3.09   0.0023 **
SandSaves              1075        759    1.42   0.1587
Scrambling             4314       2478    1.74   0.0834 .
BounceBack              568       1585    0.36   0.7203
PuttsPerRound           701      37307    0.02   0.9850
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50300 on 186 degrees of freedom
Multiple R-squared:  0.41,      Adjusted R-squared:  0.381
F-statistic: 14.3 on 9 and 186 DF,  p-value: <2e-16
```

5 cross-validation

```
> # 5 cross-validation
> cv1 <- cv.lm(data = tour, 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 1
Observations in test set: 39
                4     8     14     15     17     19     25     39     41    43     48     51    54     59    61     63
Predicted   57700 44122  91828  69736  13822  49607  88575  28224  72067 28480  17971  98677 20017  96686 33054  69446
cvpred      54890 50353  74281  66896  36506  50188  84411  32214  62708 26377  23410  77209 18307  90466 32408  50556
PrizeMoney  17516 57273  49640  53610  11989  28658  33471   8734  45752 31371  13262 132327 13865  57092 54477 217748
CV residual -37374  6920 -24641 -13286 -24517 -21530 -50940 -23480 -16956  4994 -10148  55118 -4442 -33374 22069 167192
               69     81     92     94     96     98    102    107    123    130    144    152    156    163    165    167
Predicted   18233  16663  44448  38819 21912  53145 150896 77364 43674  80585  41375 -5217  59716  63159  41089  73582
cvpred      23994  15889  36474  43265 13240  47591 120659 65417 25559  72446  38387 16770  57514  66448  50696  80956
PrizeMoney  15840   5265 100398  27673  9149  15964  70421 91406 41390  56693  24379 10715  36428  56305  19997  27657
```

```
CV residual -8154 -10624  63924 -15592 -4091 -31627 -30238 23921 13811 -15753 -14008 -6055 -21086 -10143 -30699 -53299
              171   176  177    178   180   186   187
Predicted    50479 43474 6197 244805  3895 81152 3089
cvpred       50289 43769 3712 179654  6792 68555 43384
PrizeMoney   36289 36861 9062 662771 65783 72623 90824
CV residual -14000 -6908 5350 483117 58991  4068 47440

Sum of squares = 2.93e+11    Mean square = 7.5e+09    n = 39

fold 2
Observations in test set: 40
                7     12    13     18     20     26    30     34    37     38    44     47     50     52     58    60
Predicted   11825 55613 37311 -12909  41244  79890 48019  56452 81719  41846 34538  82755  47334  78960  73474 26396
cvpred      10270 43651 33847 -33139  35679  74962 37144  55514 65799  40303 35788  73335  42912  70706  59981 11179
PrizeMoney  50620 44080 47172  20911  19683  33782 94571  37735 59151  18345 38275  10504  15187 119444 129234 45904
CV residual 40350   429 13325  54050 -15996 -41180 57427 -17779 -6648 -21958  2487 -62831 -27725  48738  69253 34725
               70    85    87    89    93    97   109   110   113   116   117   121   128   134   153   159
Predicted   -75857  63490  77199 78657  84755 -49849  45399  77700 -38735 93025 117218   5508  25243 36284  88471 71052
cvpred      -80192  55914  70660 63369  76007 -57795  39976  70638 -37014 89578 105729 -14940  25358 35436  75356 66531
PrizeMoney    2240  20612  56058 54513  37004   2692  26899  25918  12110 83483 176523  11315   5285 26532 119240 69173
CV residual  82432 -35302 -14602 -8856 -39003  60487 -13077 -44720  49124 -6095  70794  26255 -20073 -8904  43884  2642
              166   168   173   174   182   185   188   192
Predicted    82844  69826  97178  81593 15614  -4255 123151 135394
cvpred       66219  67277  83464  65867  8233 -15868 111642 120332
PrizeMoney  114055  15012 105997 150889 11187  84604 160175 170460
CV residual  47836 -52265  22533  85022  2954 100472  48533  50128

Sum of squares = 7.79e+10    Mean square = 1.95e+09    n = 40

fold 3
Observations in test set: 39
                2     9     10     28     33     36     42     53     62     64     65     66     68     73     74     77
Predicted   109843 13118  68000  96004 41968 27329  32430 74593 45278 12484  -8850  92006 38869  58704 63945  67451
cvpred       93685  4272  65844 104216 43601 27162  31563 66753 38033  8139 -17956  99646 21201  55301 64071  75007
PrizeMoney  262045 86782  23396  37751 51770 50249  14499 73819 43820  5402  10528  54862 39356 103594 57216  36918
CV residual 168360 82510 -42448 -66465  8169 23087 -17064  7066  5787 -2737  28484 -44784 18155  48293 -6855 -38089
               78    104   105    106    108    111    115    132    135    137    138    140    146    148    150    151
Predicted    24670  69750   401 114430 35264 21735 -11943  93067 82468  7478 37696  33220 79914 23286  59281 -7588
cvpred       32165  58046 -8316 103612 23077  6486 -11436 103292 78990  3740 30415  27639 60421 14221  46147 -6152
PrizeMoney    7583 117801 30068  58189 37214 42589   3025  42890 89312 11376 23403  14527 68345 16455 111028  4667
CV residual -24582  59755 38384 -45423 14137 36103  14461 -60402 10322  7636 -7012 -13112  7924  2234  64881 10819
               157   160    169    172    175    184    193
Predicted    70329 52049  70486 135285  39393 -70939  37213
cvpred       62534 53581  59172 112115  36954 -69030  32006
PrizeMoney   32843 47046  42958 106577  15098   6117  12803
CV residual -29691 -6535 -16214  -5538 -21856  75147 -19203

Sum of squares = 7.49e+10    Mean square = 1.92e+09    n = 39

fold 4
Observations in test set: 39
                1      5     11     21     35    49     55     56     57     71    72    75    79     80     83    84
Predicted   21078  42826  63621 114793  90408 56425  73186  65744  58779  69422  8780 57800 65729  28062  91833  1597
cvpred      13037  46743  70553 131048  94830 62257  84396  78714  54380  79834  9867 62021 67308  37185 110741 -2739
PrizeMoney  60661  16683  29567  79316  38455 65174  26301  22340  43951  38188 13031 82196 57824  24724  27361 55014
CV residual 47624 -30060 -40986 -51732 -56375  2917 -58095 -56374 -10429 -41646  3164 20175 -9484 -12461 -83380 57753
               86    88     90     91     95    100    120    122    124    125    129    133    136    141    147    149
Predicted    70082  76346 135248  20516  42711  96722 27634  95738  55703  50009  97371 11856 45270  7924  31552  63647
cvpred       84201  91913 133889  28002  44556 107819 35389 109268  55452  64942 107756 -3555 46763 15738  27317  66608
PrizeMoney   43173  19594 300555   7331  29296  58953 26123  18513  22467   7490  78489 25135 37869 38046  14558  19200
CV residual -41028 -72319 166666 -20671 -15260 -48866 -9266 -90755 -32985 -57452 -29267 28690 -8894 22308 -12759 -47408
              155   158   162   179    189   190    194
Predicted   47100 18015  43455 81659  55528 11778  53396
cvpred      44162 19355  38359 88552  69046 14950  61066
PrizeMoney  51005 19973  20502 89770  55581 10354  30344
CV residual  6843   618 -17857  1218 -13465 -4596 -30722

Sum of squares = 8.67e+10    Mean square = 2.22e+09    n = 39


fold 5
Observations in test set: 39
                3      6      16      22     23      24      27      29      31      32     40      45     46      67
Predicted   -24357  101255   81342   97559 10061   45069   44146   67591   50721   85132  23442  106678  19816   76572
cvpred      -79613  352822  280641  338058 31512  151986  154291  234287  180405  293376  67770  363670  67441  258476
PrizeMoney    3635  107294   26129  120927 24814   27224   20322   60073   15668  112443  56873   46377  16630   30656
CV residual  83248 -245528 -254512 -217131 -6698 -124762 -133969 -174214 -164737 -180933 -10897 -317293 -50811 -227820
               76     82     99    101    103    112    114    118    119    126    127    131    139    142
Predicted    5735   55846  42468 -15440  33505   58588   47519   69393  -38757   80849 -4762  1870  24663  145751
cvpred      17059  197488 150092 -53877 117813  202308  168439  234311 -134111  272964 -8955 13164  80436  503399
PrizeMoney  25804   16927  53530   2426  18085   18494   18721   20188    5777   18838  4444  8272  37100  224027
CV residual  8745 -180561 -96562  56303 -99728 -183814 -149718 -214123  139888 -254126 13399 -4892 -43336 -279372
              143    145    154    161    164    170    181    183    187    191    195
Predicted     79145   52949  -35509   78381   75298  1985  15880  13250  18116   99638   58089
cvpred       269610  179719 -119819  273049  254634  4276  49135  43766  67048  343902  193205
PrizeMoney   145414   53634    3816   91808   38471 11421  20064  11309  14098   68613   38043
CV residual -124196 -126085  123635 -181241 -216163  7145 -29071 -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
```



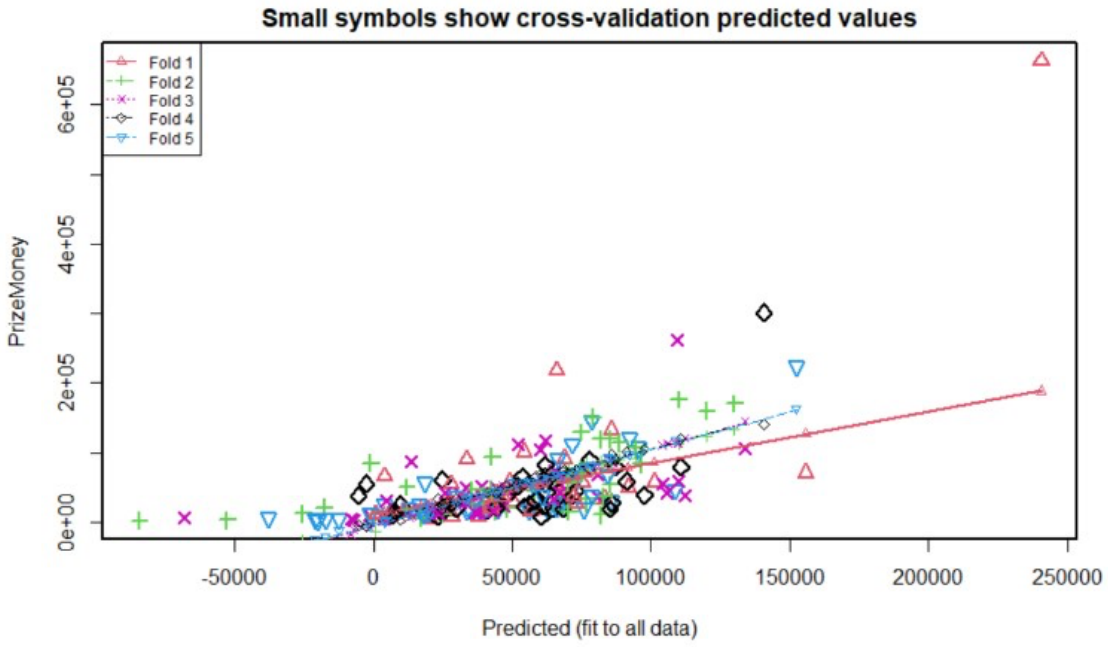
First order model after removing variables AveDrivingDistance, PuttingAverage, SandSaves, BounceBack, PuttsPerRound

```
> # First order model
> model1 <- lm(PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + Scrambling, data = tour)
> summary(model1)

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

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

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1094997     109585   -9.99  < 2e-16 ***
DrivingAccuracy    -1964        816   -2.41    0.017 *
GIR                 9743       1466    6.65  3.1e-10 ***
BirdieConversion   10670       1704    6.26  2.4e-09 ***
Scrambling          5670       1239    4.57  8.6e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50100 on 191 degrees of freedom
Multiple R-squared:  0.398,     Adjusted R-squared:  0.386
F-statistic: 31.6 on 4 and 191 DF,  p-value: <2e-16
```

Doing 5 cross-validation

```
> # 5 cross-validation
> cv2 <- cv.lm(data = tour, 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 1
Observations in test set: 39
                4     8    14    15    17    19    25    39    41    43    48    51    54    59    61    63
Predicted   57998 49185 91689 69850 18843 50321 83467 37751 67480 31646 16883 85483 10869 101099 27959 66073
cvpred      53186 49285 78911 63033 24666 46859 76473 36662 60150 35000 23480 75805 22238 86551 33528 61679
PrizeMoney  17516 57273 49640 53610 11989 28658 33471  8734 45752 31371 13262 132327 13865 57092 54477 217748
CV residual -35670  7988 -29271 -9423 -12677 -18201 -43002 -27928 -14398 -3629 -10218 56522 -8373 -29459 20949 156069
               69    81    92    94    96    98   102   107   123   130   144   152   156   163   165   167
Predicted   25769 20427 53991 37606 28345 56402 155850 68709 44477 75144 42159 777 72008 60837 42448 73004
cvpred      30274 26297 50989 36957 31942 54092 126808 61428 42879 68959 41355 10950 65522 54902 43025 65590
PrizeMoney  15840  5265 100398 27673  9149 15964 70421 91406 41390 56693 24379 10715 36428 56305 19997 27657
CV residual -14434 -21032 49409 -9284 -22793 -38128 -56387 29978 -1489 -12266 -16976 -235 -29094 1403 -23028 -37933
              171   176   177   178   180   186   196
Predicted   46676 41010 5883 240598 3729 75922 33277
cvpred      45535 40480 14789 188386 13602 68761 35730
PrizeMoney  36289 36861  9062 662771 65783 72623 90824
CV residual -9246 -3619 -5727 474385 52181 3862 55094

Sum of squares = 2.78e+11    Mean square = 7.13e+09    n = 39

fold 2
Observations in test set: 40
                7    12    13    18    20    26    30    34    37    38    44    47    50    52    58    60
Predicted   11849 58116 42605 -17854 48164 76718 42583 69123 72507 39419 39605 82019 43951 85259 74631 35557
cvpred      17706 57299 45437 -33431 47988 87017 36299 71057 73751 39282 44229 88590 45603 85735 69091 27553
PrizeMoney  50620 44080 47172 20911 19683 33782 94571 37735 59151 18345 38275 10504 15187 119444 129234 45904
CV residual 32914 -13219 1735 54342 -28305 -53235 58272 -33322 -14600 -20937 -5954 -78086 -30416 33709 60143 18351
               70    85    87    89    93    97   109   110   113   116   117   121   128   134   153
Predicted    -84595 65542 70314 85392 82943 -52882 41437 76838 -25647 96415 109904 612 16786 42643 81690
cvpred      -102978 61003 75430 83378 83471 -68041 43477 79284 -26465 100572 113053 -13396 12398 47195 76386
PrizeMoney    2240 20612 56058 54513 37004 2692 26899 25918 12110 83483 176523 11315 5285 26532 119240
CV residual 105218 -40391 -19372 -28865 -46467 70733 -16578 -53366 38575 -17089 63470 24711 -7113 -20663 42854
              159   166   168   173   174   182   185   188   192
Predicted   74201 88663 70229 94383 78885 25025 -1091 119957 130010
cvpred      81819 90094 78483 92166 75012 23885 -6453 124322 132371
PrizeMoney  69173 114055 15012 105997 150889 11187 84604 160175 170460
CV residual -12646 23961 -63471 13831 75877 -12698 91057 35853 38089

Sum of squares = 8e+10    Mean square = 2e+09    n = 40

fold 3
Observations in test set: 39
                2     9    10    28    33    36    42    53    62    64    65    66    68    73    74    77    78
Predicted   109477 13705 66067 112118 38853 33458 37938 65131 48248 3574 -1617 104012 31593 60155 66406 80578 20293
cvpred      113634 9005 68682 121281 39792 24737 36279 67589 48432 3498 -7192 110315 23471 60505 65566 82387 20711
PrizeMoney  262045 86782 23396 37751 51770 50249 14499 73819 43820 5402 10528 54862 39356 103594 57216 36918 7583
CV residual 148411 77777 -45286 -83530 11978 25512 -21780 6230 -4612 1904 17720 -55453 15885 43089 -8350 -45469 -13128
              104   105   106   108   111   115   132   135   137   138   140   146   148   150   151   157
Predicted    62047 4330 109742 32159 25631 -7300 105865 85614 22087 46728 36641 80586 12735 52253 -8368 64585
cvpred       64209 784 113673 25296 18555 -10041 113487 89435 20896 52260 35617 78427 6055 54309 -18133 70189
PrizeMoney  117801 30068 58189 37214 42589 3025 42890 89312 11376 23403 14527 68345 16455 111028 4667 32843
CV residual  53592 29284 -55484 11918 24034 13066 -70597 -123 -9520 -28857 -21090 -10082 10400 56719 22800 -37346
              160   169   172   175   184   193
Predicted   44337 67234 133847 43164 -68260 23592
cvpred      41953 70257 145158 43822 -78614 16185
PrizeMoney  47046 42958 106577 15098 6117 12803
CV residual  5093 -27299 -38581 -28724 84731 -3382

Sum of squares = 7.59e+10    Mean square = 1.95e+09    n = 39

fold 4
Observations in test set: 39
                1     5    11    21    35    49    55    56    57    71    72    75    79    80    83    84
Predicted   24623 41197 68817 110805 97604 53621 55994 54203 72816 71541 7933 61853 64600 26496 85878 -2707
cvpred      18586 42245 74656 120513 105161 52424 61159 54586 74709 71851 10823 72183 65486 28922 97634 -5903
PrizeMoney  60661 16683 29567 79316 38455 65174 26301 22340 43951 38188 13031 82196 57824 24724 27361 55014
CV residual 42075 -25562 -45089 -41197 -66706 12750 -34858 -32246 -30758 -33663 2208 10013 -7662 -4198 -70273 60917
              86    88    90    91    95   100   120   122   124   125   129   133   136   141   147   149
Predicted    62302 64625 140652 23011 44905 91103 27883 85228 61248 60388 87301 9653 47704 -5594 24678 68249
cvpred       63317 63554 140137 33543 43687 101448 23735 87695 66462 69371 89882 2483 52800 -4229 26169 74243
PrizeMoney   43173 19594 300555 7331 29296 58953 26123 18513 22467 7490 78489 25135 37869 38046 14558 19200
CV residual -20144 -43960 160418 -26212 -14391 -42495 2388 -69182 -43995 -61881 -11393 22652 -14931 42275 -11611 -55043
              155   158   162   179   189   190   194
Predicted   51641 29782 44175 77741 61704 6840 58267
cvpred      47579 33525 40841 76893 69817 9500 65581
PrizeMoney  51005 19973 20502 89770 55581 10354 30344
CV residual  3426 -13552 -20339 12877 -14236 854 -35237

Sum of squares = 7.33e+10    Mean square = 1.88e+09    n = 39
```

```
fold 5
Observations in test set: 39
                 3      6     16     22     23     24     27     29     31     32     40     45     46     67     76     82
Predicted   -12615  95669  67755  92159 16635  44308  40707  73159  38506  71536 18275 108853  33406  85467   4012  56013
cvpred      -10808  95726  65613  97721 10928  44698  42947  75646  36402  69503  9888 118549  33953  92210    712  57418
PrizeMoney    3635 107294  26129 120927 24814  27224  20322  60073  15668 112443 56873  46377  16630  30656 25804  16927
CV residual  14443  11568 -39484  23206 13886 -17474 -22625 -15573 -20734  42940 46985 -72172 -17323 -61554 25092 -40491
                99    101    103    112    114    118    119    126    127  131    139    142    143    145    154   161
Predicted   51334 -19957  44849  62841  34304  65781 -38049  75724 -17537 4170 29749 152277  78470 57312 -21267 66430
cvpred      55129 -27603  39742  62088  37224  69136 -44782  80774 -21112 2516 28218 162993  80257 57393 -24769 68989
PrizeMoney  53530   2426  18085  18494  18721  20188   5777  18838   4444 8272 37100 224027 145414 53634   3816 91808
CV residual -1599  30029 -21657 -43594 -18503 -48948  50559 -61936  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. (10 points) Evaluate scatterplots to determine which second-order terms should be tested. Test them using 5-fold cross validation and add them one-by-one until you arrive at a model you feel is appropriate. Present the model.

Solution: Final second-order model

```
> model2 <- lm(PrizeMoney ~ GIR + BirdieConversion + SandSaves + ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DAG + GPA + GBC + PA
SB + PABB + SSSB + SBBB, data = tour)
> summary(model2)

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

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
```

tour$ADD2 <- tour$AveDrivingDistance^2 tour$DA2 <-

tour$DrivingAccuracy^2 tour$G2 <- tour$GIR^2 tour$BB2 <-

tour$BounceBack^2 tour$ADDBC <- tour$AveDrivingDistance *

tour$BirdieConversion tour$ADDSS <- tour$AveDrivingDistance *

tour$SandSaves tour$DAG <- tour$DrivingAccuracy * tour$GIR

tour$GPA <- tour$GIR * tour$PuttingAverage tour$GBC <-

tour$GIR * tour$BirdieConversion tour$PASB <-

tour$PuttingAverage * tour$Scrambling tour$PABB <-

tour$PuttingAverage * tour$BounceBack tour$SSSB <-

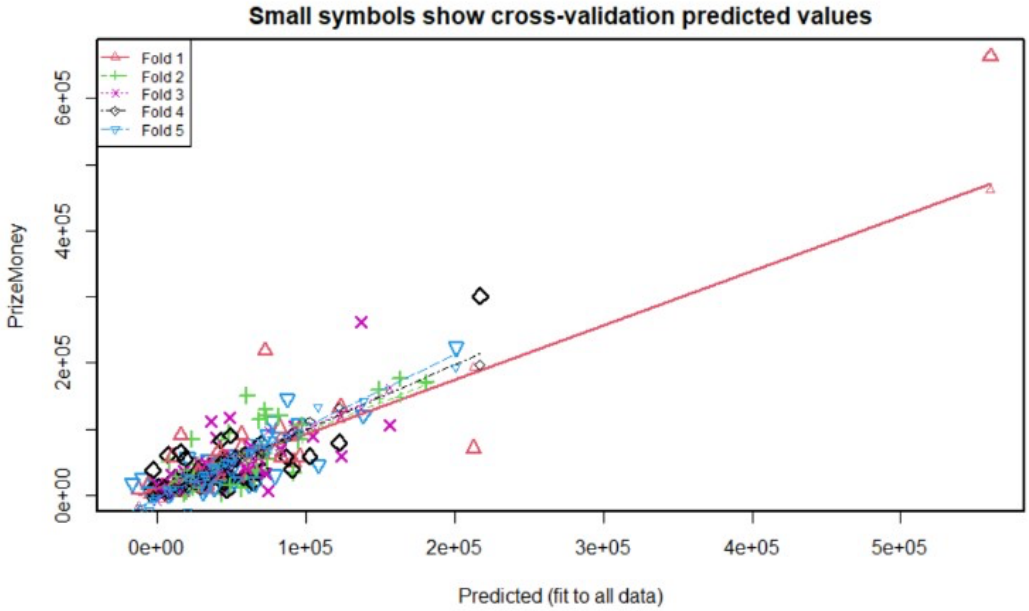
tour$SandSaves * tour$Scrambling tour$SBBB <- tour$Scrambling

* tour$BounceBack

5 cross-validation

```
> cv3 <- cv.lm(data = tour, 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 1
Observations in test set: 39
                   4     8    14    15    17    19    25    39    41    43    48    51    54    59    61    63
Predicted      59670 10045 91577 46838 26284 29705 60120 -2658 64247 30950 32550 122778  2898 95473 42213 72553
cvpred         63416 16536 86073 53717 38751 34792 66703  7869 64991 34419 26997 115005 -5186 100180 29045 60090
PrizeMoney     17516 57273 49640 53610 11989 28658 33471  8734 45752 31371 13262 132327 13865 57092 54477 217748
CV residual   -45900 40737 -36433  -107 -26762 -6134 -33232   865 -19239 -3048 -13735  17322 19051 -43088 25432 157658
                  69    81    92    94    96    98   102   107   123   130   144   152   156   163   165   167
Predicted     -6150 16489 82257 60589 -12538 32647 212655 56315 38121 82838 25115 32992 27262 46391 25935 34994
cvpred          -51 19280 61530 63969 -16730 33685 193751 57621 38749 87656 28718 35702 35394 54622 30511 52569
PrizeMoney    15840  5265 100398 27673  9149 15964  70421 91406 41390 56693 24379 10715 36428 56305 19997 27657
CV residual   15891 -14015 38868 -36296 25879 -17721 -123330 33785 2641 -30963 -4339 -24987 1034 1683 -10514 -24912
                 171   176   177   178   180   186   196
Predicted     38633 40230 12125 560222 41002 56577 15401
cvpred        46072 39696 17711 461919 41604 60060 23948
PrizeMoney    36289 36861  9062 662771 65783 72623 90824
CV residual   -9783 -2835 -8649 200852 24179 12563 66876

Sum of squares = 1.04e+11    Mean square = 2.67e+09    n = 39



fold 2
Observations in test set: 40
                   7    12    13    18    20    26    30    34    37    38    44    47    50    52    58    60    70
Predicted      7639 34128 40974 12400 26197  91537 47965 19316 56383  49138 37365  56760  27432  73358  72284 26910  43178
cvpred        11750 28588 38457 27402 29001 105330 47204 22951 53410  55292 40924  63713  30015  70722  62379 11774  83385
PrizeMoney    50620 44080 47172 20911 19683  33782 94571 37735 59151  18345 38275  10504  15187 119444 129234 45904   2240
CV residual   38870 15492  8715 -6491 -9318 -71548 47367 14784  5741 -36947 -2649 -53209 -14828  48722  66855 34130 -81145
                 85    87    89    93    97   109   110   113   116   117   121   128   134   153   159   166
Predicted      59583 47804 74343  67504  17687 24988  65792  23930  94987 163139  12745  20002  41901  81780 70228  68308
cvpred         55552 52412 63887  63436  37133 30824  67298  32614  97111 147893  21788  31190  48190  69344 74934  63407
PrizeMoney     20612 56058 54513  37004   2692 26899  25918  12110  83483 176523  11315   5285  26532 119240 69173 114055
CV residual   -34940  3646 -9374 -26432 -34441 -3925 -41380 -20504 -13628  28630 -10473 -25905 -21658  49896 -5761  50648
                168   173   174   182   185   188   192
Predicted      48677  97633  60042  13228 22804 148798 180674
cvpred         55960  86443  50472  21267  6778 146744 168965
PrizeMoney     15012 105997 150889  11187 84604 160175 170460
CV residual   -40948  19554 100417 -10080 77826  13431   1495

Sum of squares = 5.98e+10    Mean square = 1.49e+09    n = 40

fold 3
Observations in test set: 39
                   2     9    10    28    33    36    42    53    62    64    65    66    68    73    74    77    78
Predicted    136781 39482  53015  68089 27374  47387 14495 62913 28879 5782  35326 49817 42009  91900 67670  60242  46039
cvpred       130006 40435  51925  73345 25932  64787 19121 59473 26102 4927  40076 51119 33979  93207 64190  72162  54069
PrizeMoney   262045 86782  23396  37751 51770  50249 14499 73819 43820 5402  10528 54862 39356 103594 57216  36918   7583
CV residual  132039 46347 -28529 -35594 25838 -14538 -4622 14346 17718  475 -29548  3743  5377  10387 -6974 -35244 -46486
                104   105   106   108   111   115   132   135   137   138   140   146   148   150   151   157
Predicted      48597  9298 123365 17774 34443   636  59631 103999  5001  31672  29514 82364  -292  35890  5776  72482
cvpred         41352  5409 120231  8933 32966  8713  63960 104868  8216  37910  29138 72386 -9294  30267 15717  69752
PrizeMoney    117801 30068  58189 37214 42589  3025  42890  89312 11376  23403  14527 68345 16455 111028  4667  32843
CV residual    76449 24659 -62042 28281  9623 -5688 -21070 -15556  3160 -14507 -14611 -4041 25749  80761 -11050 -36909
                160   169   172   175   184   193
Predicted      32416  58498 156312 25946  74183 14435
cvpred         29220  53745 162614 24571  97313  9536
PrizeMoney     47046  42958 106577 15098   6117 12803
CV residual    17826 -10787 -56037 -9473 -91196  3267

Sum of squares = 6.06e+10    Mean square = 1.55e+09    n = 39

fold 4
Observations in test set: 39
                   1     5    11    21    35    49    55    56    57    71    72    75    79    80    83    84
Predicted      6820 21585  47727 122037  90355 15321  58753  47789 35290  70298   80.9 42195  86496 11068  58438 18439
cvpred          525 19488  51905 132452  94906  8918  66077  48757 29113  63792  507.2 52213  85136 11162  66512 14535
PrizeMoney    60661 16683  29567  79316  38455 65174  26301  22340 43951  38188 13031.0 82196  57824 24724  27361 55014
```

```
cv residual 60136 -2805 -22338 -53136 -56451 56256 -39776 -26417 14838 -25604 12523.8 29983 -27312 13562 -39151 40479
              86     88     90     91     95    100    120    122    124    125    129    133    136    141    147    149
Predicted   34329  34631 216938  46712 30229 102279 20034  64115  32497 -4176 69371  45539 27457 -3210  28689  44556
cvpred      30775  32014 197219  56200 28709 111660 11182  61455  34003 -4170 66294  46830 29778  -560  30174  53305
PrizeMoney  43173  19594 300555   7331 29296  58953 26123  18513  22467  7490 78489  25135 37869 38046  14558  19200
cv residual 12398 -12420 103336 -48869   587 -52707 14941 -42942 -11536 11660 12195 -21695  8091 38606 -15616 -34105
              155    158    162    179    189    190    194
Predicted   44192 14358 27309 48928 52118  4565 33344
cvpred      40995  5212 21200 42075 53614  -459 36012
PrizeMoney  51005 19973 20502 89770 55581 10354 30344
cv residual 10010 14761  -698 47695  1967 10813 -5668

Sum of squares = 4.64e+10    Mean square = 1.19e+09    n = 39

fold 5
Observations in test set: 39
                3     6    16    22    23    24    27    29    31    32    40    45    46    67    76    82
Predicted   -2314  95059  63431 138588 -10547  62842 24542 49806  45593  76894  20108 108018 -1000  79003 14177  61490
cvpred      -7244  92875  69197 141531 -18360  63349 25095 46428  51181  83765 -24568 132599 -3795  92282  9861  65115
PrizeMoney   3635 107294  26129 120927  24814  27224 20322 60073  15668 112443  56873  46377 16630  30656 25804  16927
cv residual 10879  14419 -43068 -20604  43174 -36125 -4773 13645 -35513  28678  81441 -86222 20425 -61626 15943 -48188
               99   101   103   112   114   118   119   126   127   131   139   142   143   145   154   161
Predicted   33877  7153 -17156  47196 18577  51062  30520  65515  8915 5914 37205 200661  86930 41904  -5502 74202
cvpred      30558  4676 -39960  49111 27468  60386  31339  80999  9844 8949 38126 193753  88540 43404 -23614 66838
PrizeMoney  53530  2426  18085  18494 18721  20188   5777  18838  4444 8272 37100 224027 145414 53634   3816 91808
cv residual 22972 -2250  58045 -30617 -8747 -40198 -25562 -62161 -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. (10 points) Beginning from scratch, engineer all possible second-order terms and add them to your dataset. From this dataset, produce a model using backward selection. Evaluate this model using 5-fold cross validation. Do you arrive at the same model as above? Explain.

Solution:

The model produced using backward selection has one extra variable 'DrivingAccuracy'.

Second-order model by backward selection

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

Initial Model:
PrizeMoney ~ AveDrivingDistance + DrivingAccuracy + GIR + PuttingAverage +
    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
>
```

```
> model4 <- lm(PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + SandSaves + ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DA
G + GPA + GBC + PASB + PABB + SSSB + SBBB, data = tour)
> summary(model4)

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

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
```

tour$ADD2 <- tour$AveDrivingDistance^2
tour$DA2 <- tour$DrivingAccuracy^2 tour$G2 <- tour$GIR^2

tour$BB2 <- tour$BounceBack^2 tour$ADDBC <-

tour$AveDrivingDistance * tour$BirdieConversion tour$ADDSS <-

tour$AveDrivingDistance * tour$SandSaves tour$DAG <-

tour$DrivingAccuracy * tour$GIR tour$GPA <- tour$GIR *

tour$PuttingAverage tour$GBC <- tour$GIR *

tour$BirdieConversion tour$PASB <- tour$PuttingAverage *

tour$Scrambling tour$PABB <- tour$PuttingAverage *

tour$BounceBack tour$SSSB <- tour$SandSaves *

tour$Scrambling tour$SBBB <- tour$Scrambling *

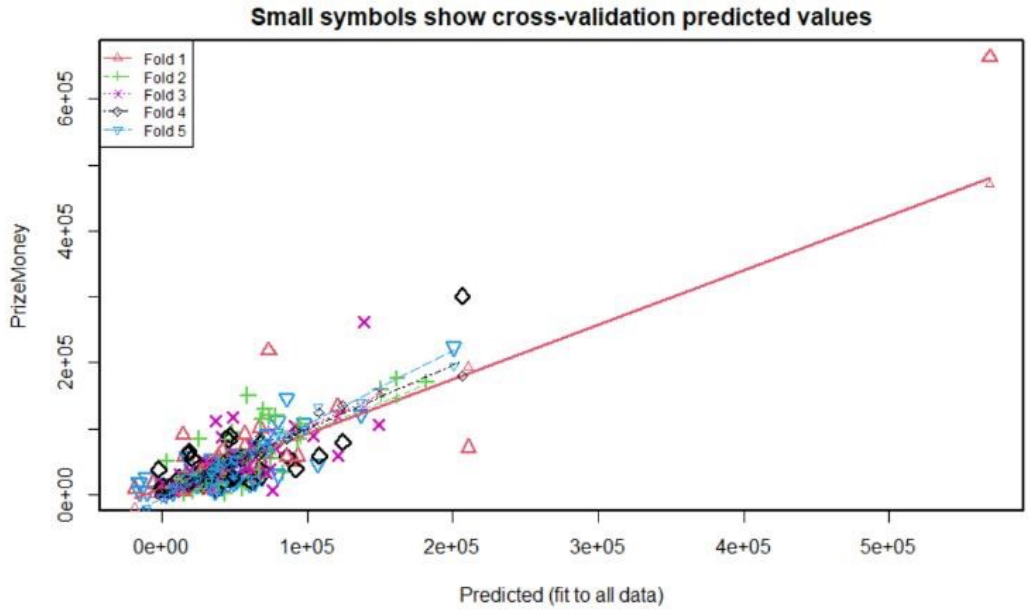
tour$BounceBack

5cross validation

```
> cv4 <- cv.lm(data = tour, 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 1
Observations in test set: 39
                4     8     14    15     17    19     25   39     41    43    48     51    54     59    61     63
Predicted   59437 15755  91180 46245  26524 32813  65962  928  64769 33726 25319 121123 15352  93221 38853  73539
cvpred      63185 18720  86108 53420  38519 35958  70624 9233  65154 35432 23023 114975  1388  99407 27142  60641
PrizeMoney  17516 57273  49640 53610  11989 28658  33471 8734  45752 31371 13262 132327 13865  57092 54477 217748
CV residual -45669 38553 -36468   190 -26530 -7300 -37153 -499 -19402 -4061 -9761  17352 12477 -42315 27335 157107
                69     81     92     94     96     98    102   107   123    130   144    152   156   163   165    167
Predicted   -5576.8  15746  67577  62350 -18746  30917  210563 57000 40712  85802 25685  32660 27996 48492 22915  33280
cvpred         15.6  18950  54209  64483 -20158  32903  194068 57977 39952  89617 28783  34937 35471 55369 28507  51599
PrizeMoney  15840.0   5265 100398  27673   9149  15964   70421 91406 41390  56693 24379  10715 36428 56305 19997  27657
CV residual 15824.4 -13685  46189 -36810  29307 -16939 -123647 33429  1438 -32924 -4404 -24222   957   936 -8510 -23942
              171   176   177    178   180   186   196
Predicted   39502 37264 13957 568943 41301 57185 14451
cvpred      46361 38273 18242 470356 41703 60629 22826
PrizeMoney  36289 36861  9062 662771 65783 72623 90824
CV residual -10072 -1412 -9180 192415 24080 11994 67998

Sum of squares = 1.02e+11    Mean square = 2.6e+09    n = 39
```

```
fold 2
Observations in test set: 40
                7     12     13     18     20     26     30     34     37     38     44     47     50     52     58     60     70
Predicted    3086 34097 38678 11216 26569  83232 48442 21900 55403  57936 35551  55328  27477  74330  70103 29787  43287
cvpred       5129 28016 35894 24821 29026  95479 47258 25291 51950  63899 38909  60429  29904  72081  60278 16079  84040
PrizeMoney  50620 44080 47172 20911 19683  33782 94571 37735 59151  18345 38275  10504  15187 119444 129234 45904   2240
CV residual 45491 16064 11278 -3910 -9343 -61697 47313 12444  7201 -45554  -634 -49925 -14717  47363  68956 29825 -81800
               85     87     89     93     97    109    110    113    116    117    121    128    134    153    159    166
Predicted    58730 44105 74846  67003  15571 26125  66946  27717  93920 161546  9088  21296  38915  78616 66808  67910
cvpred       54744 47912 64508  62745  35678 32127  67235  36225  96062 145797 17590  31736  44205  66154 71528  62971
PrizeMoney   20612 56058 54513  37004   2692 26899  25918  12110  83483 176523 11315   5285  26532 119240 69173 114055
CV residual -34132  8146 -9995 -25741 -32986 -5228 -41317 -24115 -12579  30726 -6275 -26451 -17673  53086 -2355  51084
              168    173    174    182    185    188    192
Predicted    39419  95469  59008 13694 25583 150134 181591
cvpred       44015  84161  49443 21009  9242 147260 169893
PrizeMoney   15012 105997 150889 11187 84604 160175 170460
CV residual -29003  21836 101446 -9822 75362  12915    567

Sum of squares = 5.87e+10    Mean square = 1.47e+09    n = 40

fold 3
Observations in test set: 39
                2      9     10     28     33     36     42     53     62     64     65     66     68     73     74     77
Predicted   139165 40901  52327  74482 28191  48366 16363 63079 29642 -1245  42142 47509 37595  91588 68709  62005
cvpred      131122 41818  51479  79528 26634  65735 20278 59649 26596  -980  45315 49696 31114  93696 65090  73973
PrizeMoney  262045 86782  23396  37751 51770  50249 14499 73819 43820  5402  10528 54862 39356 103594 57216  36918
CV residual 130923 44964 -28083 -41777 25136 -15486 -5779 14170 17224  6382 -34787  5166  8242   9898 -7874 -37055
               78    104    105    106    108    111    115    132    135    137    138    140    146    148    150    151
Predicted    34886  48561 12714 121285 19005 35499   94.9  59332 104323  6242  30395  31724 78275  1488  37076   7934
cvpred       46064  41092  8032 118368  9639 34085  7583.4  64107 105097  9480  36511  30894 69250 -7943  30453  17591
PrizeMoney    7583 117801 30068  58189 37214 42589  3025.0  42890  89312 11376  23403  14527 68345 16455 111028   4667
CV residual -38481  76709 22036 -60179 27575  8504 -4558.4 -21217 -15785  1896 -13108 -16367  -905 24398  80575 -12924
              157    160    169    172    175    184    193
Predicted    71741 31547  57813 149518 22165  75753 12784
cvpred       69355 28180  52984 156621 21366  97787  7991
PrizeMoney   32843 47046  42958 106577 15098   6117 12803
CV residual -36512 18866 -10026 -50044 -6268 -91670  4812

Sum of squares = 5.95e+10    Mean square = 1.52e+09    n = 39


fold 4
Observations in test set: 39
                1      5     11     21     35     49     55     56     57     71     72     75     79     80     83     84
Predicted   17919 23038  45523 124276  92088 18828  56257  46250 35782  67512   502 46123  86146 11536  67698 20660
cvpred      23455 21207  49046 135529  96549 14955  62071  47279 29205  60155  1243 58861  83685 13468  87369 17077
PrizeMoney  60661 16683  29567  79316  38455 65174  26301  22340 43951  38188 13031 82196  57824 24724  27361 55014
CV residual 37206 -4524 -19479 -56213 -58094 50219 -35770 -24939 14746 -21967 11788 23335 -25861 11256 -60008 37937
               86     88     90     91     95    100    120    122    124    125    129    133    136    141    147    149    155
Predicted   34376 32280 206509  37607 27659 108105 25665  61658  31958 -871 68369  48068 27702 -2220  28707  42721 48180
cvpred      29919 28480 179984  42611 25401 124593 21123  57626  33831 1284 63352  50404 30252  -111  29254  52261 47833
PrizeMoney  43173 19594 300555   7331 29296  58953 26123  18513  22467 7490 78489  25135 37869 38046  14558  19200 51005
CV residual 13254 -8886 120571 -35280  3895 -65640  5000 -39113 -11364 6206 15137 -25269  7617 38157 -14696 -33061  3172
              158    162    179    189    190    194
Predicted   19152 29885 46482 54548  5219 34673
cvpred      10703 24586 38069 56750   738 39347
PrizeMoney  19973 20502 89770 55581 10354 30344
CV residual  9270 -4084 51701 -1169  9616 -9003

Sum of squares = 4.87e+10    Mean square = 1.25e+09    n = 39

fold 5
Observations in test set: 39
                3      6     16     22     23     24     27     29     31     32     40     45     46     67     76     82
Predicted   -10358  97593  68650 136949 -10807  62869 23992 49449  44913  79533  16677 107423  -194  80206 15022  60325
cvpred      -25054  98419  79348 139822 -20652  63710 24268 45118  50843  89374 -31239 133823 -3700  94844 11989  64061
PrizeMoney    3635 107294  26129 120927  24814  27224 20322 60073  15668 112443  56873  46377 16630  30656 25804  16927
CV residual  28689   8875 -53219 -18895  45466 -36486 -3946 14955 -35175  23069  88112 -87446 20330 -64188 13815 -47134
               99    101    103    112    114    118    119    126    127    131    139    142    143    145    154    161
Predicted   34101  2912 -15654  49495 17068  51150  36517  63766  7491  8534 36045 200877  86174 41199 -15390 72508
cvpred      30079 -4627 -38980  52371 26151  61522  42748  78761  8811 12089 37974 194701  87473 43216 -46525 64620
PrizeMoney  53530  2426  18085  18494 18721  20188   5777  18838  4444  8272 37100 224027 145414 53634   3816 91808
CV residual 23451  7053  57065 -33877 -7430 -41334 -36971 -59923 -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. (10 points) You have used two procedures to build a second-order model. Compare these two procedures. Which do you think is “best”? Explain.

Solution: Although the backward selection technique is quicker than the first, it adds a driving accuracy variable with a p value greater than 0.05 that doesn't improve the model. In my opinion, the backward selection model is inferior to the first model without the additional variable. However, it might lead to overfitting. Backward selection can characterise the variation of the dependent variable.