honor statement: "I have completed this work independently. The solutions given are entirely my own work." Your submission must be submitted as a PDF.

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.

Since R squared does not offer any significant insight into how well our regression model can predict future value, we could apply cross - validation which is a process in which we obtain empirical evidence as to its capacity to make accurate predictions for new samples of data.

In cross-validation the original sample is split into two parts. One part is called a training sample, the other part is a testing sample. We will use the training data to train the model and use test data to test the model to see if we need to adjust any of the parameters or select any additional features.
Finally we are going to see how well the model performs on the validation data.
The validation error gives an unbiased estimate of predictive power of model, because in no way was validation used to construct the models.

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.

1. Build up the first model with all variables and check the F-test, F-test is good enough to reject the null hypothesis which means at least one H is not equal to zero.
2. Then check ad r-squared - 0.381
3. Then Check the p-value of all of the variables, first I would remove variable of PuttsPerRound since it p-value is too large to reject the null hypothesis, second, I would remove variable of BounceBack with p-value 0.7173 , third I would remove variable of PuttingAverage with p-value 0.63568 and last I would remove variable of AveDrivingDistance with p-value 0.3974 , then I'll have the first - order model with variables :DrivingAccuracy,GIR ,BirdieConversion,SandSaves and Scrambling

4. By evaluating the model by using 5-fold cross validation,the overall average of the mean square is 1.23e+10 which is very small, which tells us the regression line is close to a set of points.

```
> model_5 <-lm(PrizeMoney~., data = pgatour2006_4)
> summary(model_5)

Call:
lm(formula = PrizeMoney ~ ., data = pgatour2006_4)

Residuals:
   Min     1Q Median     3Q    Max
-80972 -26436  -6308  17398 420690

Coefficients:
                   Estimate      Std. Error    t value    Pr(>|t|)
(Intercept)        -1103461       109277       -10.10   <2e-16 ***
DrivingAccuracy       -1848          816        -2.27   0.0246 *
GIR                   10136         1481         6.84    1e-10 ***
BirdieConversion      10274         1715         5.99    1e-08 ***
SandSaves              1173          736         1.59   0.1130
Scrambling             4446         1454         3.06   0.0026 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49900 on 190 degrees of freedom
Multiple R-squared:  0.406,    Adjusted R-squared:  0.391
F-statistic:   26 on 5 and 190 DF,  p-value: <2e-16
```

```
> out<- cv.lm(data = pgatour2006_4 , form.lm = (PrizeMoney~.),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          0.65938
```

```
GIR                   1 1.54e+11   1.54e+11   61.93      2.6e-13 ***
BirdieConversion      1 1.10e+11   1.10e+11   44.28      3.0e-10 ***
SandSaves             1 3.56e+10   3.56e+10   14.29      0.00021 ***
Scrambling            1 2.32e+10   2.32e+10    9.34      0.00256 **
Residuals           190 4.73e+11   2.49e+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
Predicted   57967 43920  92082 70710 12865  51957  88566  31545  70592 33185  19623
95639
cvpred      52906 46744  79049 63480 21477  47388  79575  33000  61523 35746  24779
81078
PrizeMoney  17516 57273  49640 53610 11989  28658  33471   8734  45752 31371  13262
132327
CV residual -35390 10529 -29409 -9870 -9488 -18730 -46104 -24266 -15771 -4375 -11517
51249
             54    59    61    63    69    81    92    94    96    98   102   107
Predicted   19138  93960 31744  71362  22756  17472  50717 33927  23780  51642 151393
77723
cvpred      26826  82970 35557  64557  28689  24747  49093 34654  29508  51828 124564
65891
PrizeMoney  13865  57092 54477 217748  15840   5265 100398 27673   9149  15964  70421
91406
CV residual -12961 -25878 18920 153191 -12849 -19482  51305 -6981 -20359 -35864 -54143
25515
            123   130   144   152   156   163   165   167   171   176   177   178
Predicted   43905  73875  42321 -6655  62584 60392  43959  72021  51183 37639  5430
242081
cvpred      42321  68609  41162  6899  60836 54428  43753  65131  47744 38475 14425
189068
PrizeMoney  41390  56693  24379 10715  36428 56305  19997  27657  36289 36861  9062
662771
CV residual  -931 -11916 -16783  3816 -24408  1877 -23756 -37474 -11455 -1614 -5363
473703
            180   186   196
Predicted    9033 76700 33464
cvpred      16195 69282 35710
PrizeMoney  65783 72623 90824
CV residual 49588  3341 55114
```

Sum of squares = 2.75e+11 Mean square = 7.04e+09 n = 39

fold 2
Observations in test set: 40
	7	12	13	18	20	26	30	34	37	38	44	47
Predicted	10029	53101	40840	-23145	43791	75790	43926	64579	73789	50061	36373	83589
cvpred	11513	43932	35759	-34753	35792	73153	32289	56245	64489	42929	33320	77066
PrizeMoney	50620	44080	47172	20911	19683	33782	94571	37735	59151	18345	38275	10504
CV residual	39107	148	11413	55664	-16109	-39371	62282	-18510	-5338	-24584	4955	-66562

	50	52	58	60	70	85	87	89	93	97	109
Predicted	47529	82506	75528	28420	-75983	66131	73759	79108	83933	-48646	48389
cvpred	41977	71845	60902	17181	-81502	53388	67197	65568	72709	-55225	43581
PrizeMoney	15187	119444	129234	45904	2240	20612	56058	54513	37004	2692	26899
CV residual	-26790	47599	68332	28723	83742	-32776	-11139	-11055	-35705	57917	-16682

	110	113	116	117	121	128	134	153	159	166	168	173
Predicted	79849	-32272	95946	117747	-1033	22746	36867	84246	71006	82585	75754	99048
cvpred	70847	-32575	86546	104609	-13709	15336	32434	69069	65365	71543	71093	84493
PrizeMoney	25918	12110	83483	176523	11315	5285	26532	119240	69173	114055	15012	105997
CV residual	-44929	44685	-3063	71914	25024	-10051	-5902	50171	3808	42512	-56081	21504

	174	182	185	188	192
Predicted	83811	17562	-7332	124300	133246
cvpred	69925	11007	-12511	111921	118251
PrizeMoney	150889	11187	84604	160175	170460
CV residual	80964	180	97115	48254	52209

Sum of squares = 7.65e+10 Mean square = 1.91e+09 n = 40

fold 3
Observations in test set: 39
	2	9	10	28	33	36	42	53	62	64	65	66
Predicted	115199	11238	63580	102604	39329	25662	31980	70211	47280	10828	-7330	91684
cvpred	101898	6392	58509	103431	36139	21771	29882	60641	42104	8324	-10045	88485

PrizeMoney 262045 86782 23396 37751 51770 50249 14499 73819 43820 5402 10528
54862
CV residual 160147 80390 -35113 -65680 15631 28478 -15383 13178 1716 -2922 20573
-33623
            68   73   74   77   78   104   105   106   108   111   115   132
Predicted   40141 59816 71451 63401 20471 67837 3392 112731 38573 24309 -9940
93789
cvpred   23100 50615 63146 63700 21341 59144 -860 100394 25024 14760 -7948 92961
PrizeMoney 39356 103594 57216 36918   7583 117801 30068 58189 37214 42589 3025
42890
CV residual 16256 52979 -5930 -26782 -13758 58657 30928 -42205 12190 27829 10973
-50071
           135   137   138   140   146   148   150   151   157   160   169   172
Predicted   86101 11549 42957 36825 86064 20298 61270 -9396 67358 49420 71760
140194
cvpred   77040 12742 42083 31628 70597 8186 51798 -12566 62113 39554 61815
126191
PrizeMoney 89312 11376 23403 14527 68345 16455 111028   4667 32843 47046 42958
106577
CV residual 12272 -1366 -18680 -17101 -2252 8269 59230 17233 -29270 7492 -18857
-19614
            175   184   193
Predicted   45419 -72171 31056
cvpred   42064 -66953 21365
PrizeMoney   15098   6117 12803
CV residual -26966 73070 -8562

Sum of squares = 6.69e+10   Mean square = 1.72e+09   n = 39

fold 4
Observations in test set: 39
             1   5   11   21   35   49   55   56   57   71   72   75
Predicted   25190 45184 64792 111368 94782 57048 68552 65154 62451 71944 9719
55404
cvpred   18901 45955 65004 114278 95755 55205 75898 67016 56158 68686 13597
59502
PrizeMoney 60661 16683 29567 79316 38455 65174 26301 22340 43951 38188 13031
82196
CV residual 41760 -29272 -35437 -34962 -57300 9969 -49597 -44676 -12207 -30498 -566
22694
            79   80   83   84   86   88   90   91   95   100   120   122
Predicted   64785 30530 90385 2133 68861 74039 133083 22307 39760 96999 22220
93275

cvpred 61729 33964 98259 1869 69166 73104 120360 31954 33635 103948 14938 94104
PrizeMoney 57824 24724 27361 55014 43173 19594 300555 7331 29296 58953 26123 18513
CV residual -3905 -9240 -70898 53145 -25993 -53510 180195 -24623 -4339 -44995 11185 -75591

	124	125	129	133	136	141	147	149	155	158	162	179
Predicted	52827	49061	91406	8266	48307	3944	25887	66135	51841	29256	40284	74619
cvpred	51147	49630	90709	761	51480	9918	26826	67859	45243	31756	33298	68095
PrizeMoney	22467	7490	78489	25135	37869	38046	14558	19200	51005	19973	20502	89770
CV residual	-28680	-42140	-12220	24374	-13611	28128	-12268	-48659	5762	-11783	-12796	21675

	189	190	194
Predicted	59705	10261	50708
cvpred	63371	15134	51669
PrizeMoney	55581	10354	30344
CV residual	-7790	-4780	-21325

Sum of squares = 7.51e+10 Mean square = 1.93e+09 n = 39

fold 5
Observations in test set: 39

	3	6	16	22	23	24	27	29	31	32
Predicted	-21505	94796	80900	101267	15099	44108	46742	68831	45699	83073
cvpred	-89817	428491	369185	460450	69451	194108	214808	311622	211411	376053
PrizeMoney	3635	107294	26129	120927	24814	27224	20322	60073	15668	112443
CV residual	93452	-321197	-343056	-339523	-44637	-166884	-194486	-251549	-195743	-263610

	40	45	46	67	76	82	99	101	103	112
Predicted	20343	109565	24708	83474	1031	52215	45340	-10577	30397	62815
cvpred	81693	491619	109709	373672	2313	237252	207907	-43433	138656	285264
PrizeMoney	56873	46377	16630	30656	25804	16927	53530	2426	18085	18494
CV residual	-24820	-445242	-93079	-343016	23491	-220325	-154377	45859	-120571	-266770

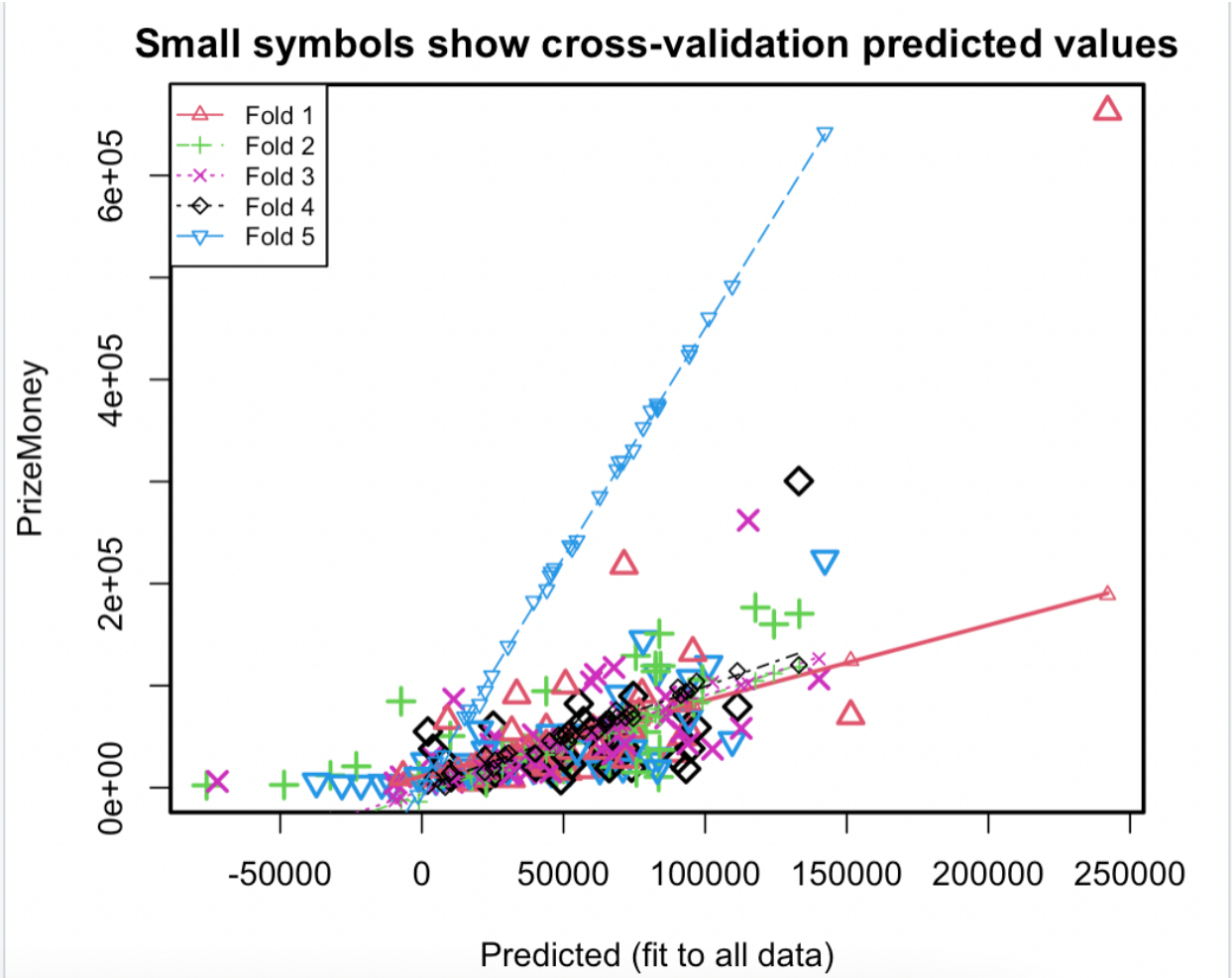
	114	118	119	126	127	131	139	142	143	145
Predicted	39449	70980	-37219	83166	-14185	4915	22372	142269	78114	52996
cvpred	182700	320060	-168912	371733	-56369	25148	94591	642078	352935	234770
PrizeMoney	18721	20188	5777	18838	4444	8272	37100	224027	145414	53634
CV residual	-163979	-299872	174689	-352895	60813	-16876	-57491	-418051	-207521	-181136

	154	161	164	170	181	183	187	191	195
Predicted	-28228	69624	74639	-1339	16543	7419	16649	94369	54737
cvpred	-120703	319360	330942	-7422	69319	31446	76141	423902	242057

PrizeMoney 3816 91808 38471 11421 20064 11309 14098 68613 38043
CV residual 124519 -227552 -292471 18843 -49255 -20137 -62043 -355289 -204014

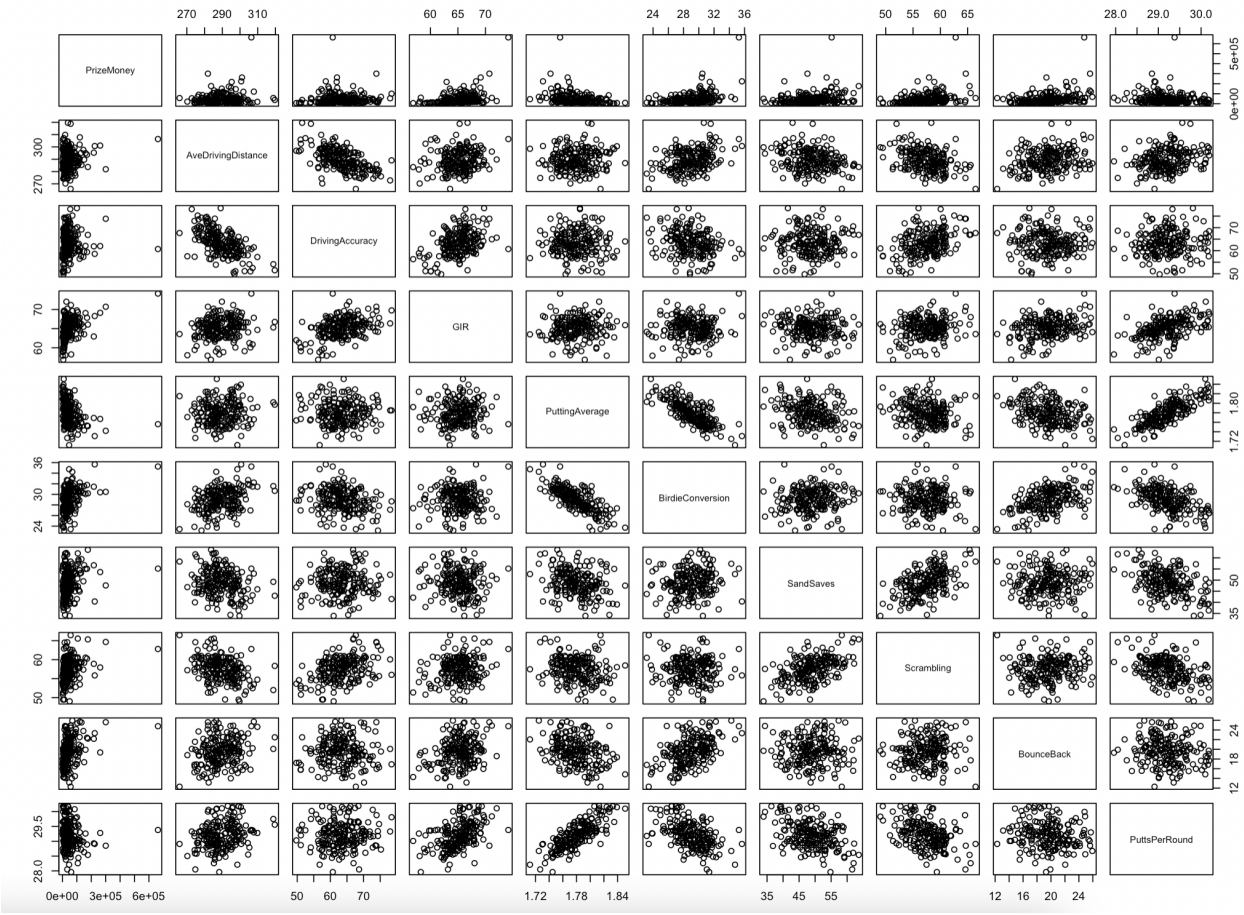
Sum of squares = 1.93e+12 Mean square = 4.94e+10 n = 39

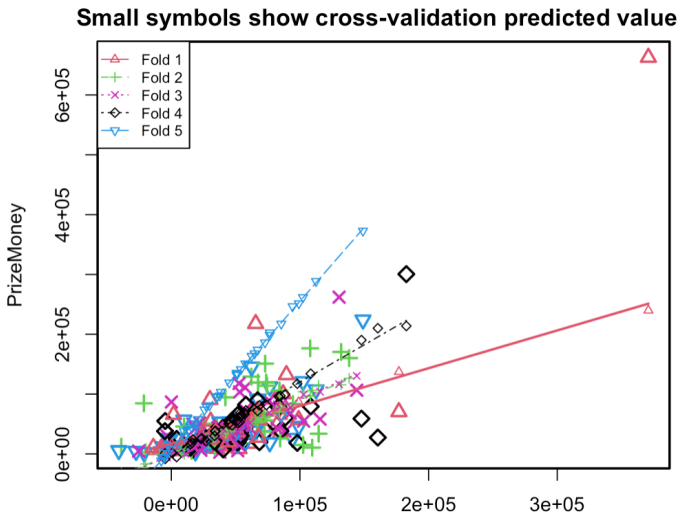
Overall (Sum over all 39 folds)
ms
1.23e+10



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.

By checking the scatterplot of all the variables to PrizeMoney, I would use GIR and BirdieConversion these two variables into my second-order terms model.
First, I would use GIR square in my second-order terms model, and the Adjusted R-squared has improved to 0.489 from 0.391. If we Evaluated model by using 5 fold cross validation, the overall average of the mean square is 4.49e+09 .
Secondly, I would use BirdieConversion square in my second-order terms model and the Adjusted R-squared has improved to 0.489 from 0.529 If we Evaluated model by using 5 fold cross validation, the overall average of the mean square is 7.27e+09 .






```
> model_sq1<-lm(PrizeMoney~., data = pgatour2006_4)
> summary(model_sq1)

Call:
lm(formula = PrizeMoney ~ ., data = pgatour2006_4)

Residuals:
   Min      1Q  Median      3Q     Max
-133180  -24211   -4754   19152  291915

Coefficients:
                  Estimate     Std. Error    t value      Pr(>|t|)
(Intercept)       5974716      1156171       5.17         6.0e-07 ***
DrivingAccuracy     -1171          755      -1.55         0.1223
GIR               -209472        35763      -5.86         2.1e-08 ***
BirdieConversion    10799         1572       6.87         9.1e-11 ***
SandSaves            1077          674       1.60         0.1119
Scrambling           4361         1331       3.28          0.0013 **
GIR_SQ               1689          275       6.15          4.6e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45700 on 189 degrees of freedom
Multiple R-squared:  0.505,    Adjusted R-squared:  0.489
F-statistic: 32.2 on 6 and 189 DF,  p-value: <2e-16

> out_1<- cv.lm(data = pgatour2006_4 , form.lm = (PrizeMoney~.),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    0.23   0.630
GIR                1 1.54e+11 1.54e+11   73.91 3.1e-15 ***
BirdieConversion   1 1.10e+11 1.10e+11   52.84 9.3e-12 ***
SandSaves          1 3.56e+10 3.56e+10   17.06 5.4e-05 ***
Scrambling         1 2.32e+10 2.32e+10   11.15   0.001 **
GIR_SQ             1 7.87e+10 7.87e+10   37.76 4.6e-09 ***
Residuals        189 3.94e+11 2.08e+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
Predicted    60434 48929  82240 62638   280  46102  84549  52662  56382 22942  41699
89477
cvpred       54695 48590  77507 61539 16439  45971  78954  40640  57822 31900  31333
80476
PrizeMoney   17516 57273  49640 53610 11989  28658  33471   8734  45752 31371  13262
132327
CV residual -37179  8683 -27867 -7929 -4450 -17313 -45483 -31906 -12070  -529 -18071
51851
              54     59    61     63    69     81     92     94     96     98    102   107
Predicted    43386  98963 30613  65581  9779   7641  86898  45299  35104  46648 176923
80988
cvpred       33302  86633 34583  63153 23600  20500  62293  38866  32912  50277 137256
67938
PrizeMoney   13865  57092 54477 217748 15840   5265 100398  27673   9149  15964  70421
91406
CV residual -19437 -29541 19894 154595 -7760 -15235  38105 -11193 -23763 -34313 -66835
23468
             123    130    144    152    156   163    165    167   171   176   177    178
Predicted    39011  68401  32942 -13998  51748 58114  44385  67680 37794 34598 -8041
370856
cvpred       41003  67468  38351   3258  58033 54508  43931  64533 43344 37806  8408
239833
PrizeMoney   41390  56693  24379  10715  36428 56305  19997  27657 36289 36861  9062
662771
CV residual   387 -10775 -13972   7457 -21605  1797 -23934 -36876 -7055  -945   654 422938
             180   186   196
Predicted    1754 66960 30012
cvpred      12268 67062 34314
```

PrizeMoney 65783 72623 90824
CV residual 53515 5561 56510

Sum of squares = 2.31e+11   Mean square = 5.92e+09   n = 39

fold 2
Observations in test set: 40
             7   12   13   18   20   26   30   34   37   38   44   47
Predicted   25141 43215 35444 42194 33151 114516 42029 66647 65181 52428 32200
109557
cvpred   32153 35530 34188 39547 27814 114650 33873 59871 57403 51179 32079
106370
PrizeMoney 50620 44080 47172 20911 19683 33782 94571 37735 59151 18345 38275
10504
CV residual 18467 8550 12984 -18636 -8131 -80868 60698 -22136 1748 -32834 6196 -95866
            50   52   58   60   70   85   87   89   93   97   109
Predicted   38565 73741 62213 9943 -9041.04 49611 72253 66695 72394 -20080 39367
cvpred   36161 61860 48466 -368   -6.23 38278 68767 52840 61919 -16737 37932
PrizeMoney   15187 119444 129234 45904 2240.00 20612 56058 54513 37004   2692
26899
CV residual -20974 57584 80768 46272 2246.23 -17666 -12709 1673 -24915 19429 -11033
            110   113   116   117   121   128   134   153   159   166   168   173
Predicted   84593 -38916 94521 107968 33916 15720 32860 67437 68145 74415 102507
84083
cvpred   78366 -31596 84272 96690 28645 14311 32809 52809 64783 62784 103984
70042
PrizeMoney   25918 12110 83483 176523 11315 5285 26532 119240 69173 114055 15012
105997
CV residual -52448 43706 -789 79833 -17330 -9026 -6277 66431 4390 51271 -88972
35955
            174   182   185   188   192
Predicted   72958 7664 -21347 138235 132045
cvpred   60301 5742 -22695 126333 115731
PrizeMoney 150889 11187 84604 160175 170460
CV residual 90588 5445 107299 33842 54729

Sum of squares = 8.77e+10   Mean square = 2.19e+09   n = 40

fold 3
Observations in test set: 39
              2   9   10   28   33   36   42   53   62   64   65   66
Predicted   130435   71.9 51011 88449 22047 16670 27551 57259 36260 51647 -7095
102821

cvpred 116465 -4745.8 46465 89080 17537 13665 28561 48602 33006 54450 -7269 99145
PrizeMoney 262045 86782.0 23396 37751 51770 50249 14499 73819 43820 5402 10528 54862
CV residual 145580 91527.8 -23069 -51329 34233 36584 -14062 25217 10814 -49048 17797 -44283
68 73 74 77 78 104 105 106 108 111 115 132
Predicted 47623 53153 54611 57247 22576 52662 -3325 115629 51961 18785 37749 82078
cvpred 30411 44828 47423 61131 26199 44240 -6195 103905 36735 10918 45666 83197
PrizeMoney 39356 103594 57216 36918 7583 117801 30068 58189 37214 42589 3025 42890
CV residual 8945 58766 9793 -24213 -18616 73561 36263 -45716 479 31671 -42641 -40307
135 137 138 140 146 148 150 151 157 160 169 172
Predicted 76076 1815 40311 23151 92497 11072.1 57342 -25366 50373 37528 61147 144146
cvpred 69064 4156 41944 17347 77801 -32.5 51118 -29331 45425 29159 51497 130397
PrizeMoney 89312 11376 23403 14527 68345 16455.0 111028 4667 32843 47046 42958 106577
CV residual 20248 7220 -18541 -2820 -9456 16487.5 59910 33998 -12582 17887 -8539 -23820
175 184 193
Predicted 52484 23006 17506
cvpred 53501 35623 9543
PrizeMoney 15098 6117 12803
CV residual -38403 -29506 3260

Sum of squares = 6.81e+10 Mean square = 1.75e+09 n = 39

fold 4
Observations in test set: 39
1 5 11 21 35 49 55 56 57 71 72 75
Predicted 88560 34085 49541 108376 84908 51661 53658 56838 53128 73906 884 57972
cvpred 99619 38245 57498 134135 96949 51235 68503 69400 53343 81520 1497 72915
PrizeMoney 60661 16683 29567 79316 38455 65174 26301 22340 43951 38188 13031 82196
CV residual -38958 -21562 -27931 -54819 -58494 13939 -42202 -47060 -9392 -43332 11534 9281
79 80 83 84 86 88 90 91 95 100 120 122

Predicted 54488 23110 160541 -4725 56914 68511 182690 16733 34102 147796 27135 97917
cvpred 64177 25049 210110 -8311 66137 81240 213831 24260 37789 190282 21353 117352
PrizeMoney 57824 24724 27361 55014 43173 19594 300555 7331 29296 58953 26123 18513
CV residual -6353 -325 -182749 63325 -22964 -61646 86724 -16929 -8493 -131329 4770 -98839
124 125 129 133 136 141 147 149 155 158 162 179
Predicted 44677 40330 83310 4146 34248 -4919 15996 51513 47215 23673 33579 67099
cvpred 49956 49948 96194 -5249 38537 1051 18664 56543 47841 26275 30117 71775
PrizeMoney 22467 7490 78489 25135 37869 38046 14558 19200 51005 19973 20502 89770
CV residual -27489 -42458 -17705 30384 -668 36995 -4106 -37343 3164 -6302 -9615 17995
189 190 194
Predicted 47199 14879 53907
cvpred 59213 15889 64385
PrizeMoney 55581 10354 30344
CV residual -3632 -5535 -34041

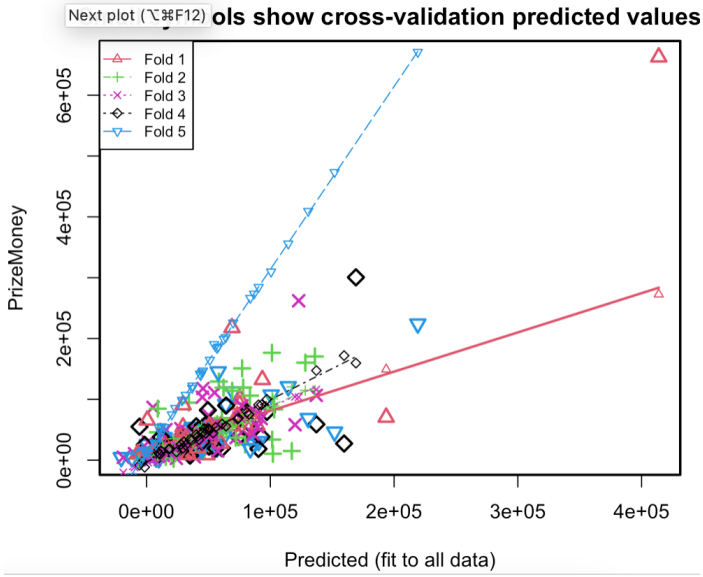
Sum of squares = 1.01e+11 Mean square = 2.58e+09 n = 39

fold 5
Observations in test set: 39
3 6 16 22 23 24 27 29 31 32 40
Predicted 18745 112520 94276 102388 -1128 35733 35274 54124 27439 76510 9717
cvpred 51070 288725 246581 261791 8001 96081 94823 141329 79175 202824 37902
PrizeMoney 3635 107294 26129 120927 24814 27224 20322 60073 15668 112443 56873
CV residual -47435 -181431 -220452 -140864 16813 -68857 -74501 -81256 -63507 -90381 18971
45 46 67 76 82 99 101 103 112 114 118
Predicted 99151 20751 72830 -2949 51435 37681 -21228 38514 50722 29665 57595
cvpred 251445 57740 186099 1223 133675 97449 -39362 103318 135246 81012 150920
PrizeMoney 46377 16630 30656 25804 16927 53530 2426 18085 18494 18721 20188
CV residual -205068 -41110 -155443 24581 -116748 -43919 41788 -85233 -116752 -62291 -130732
119 126 127 131 139 142 143 145 154 161 164
Predicted -40833 76530 -27399 -4876 15558 148971 62445 45033 -7800 67108 64765
cvpred -92053 198651 -59592 -5137 45044 372809 163622 119101 -11348 173900 167850
PrizeMoney 5777 18838 4444 8272 37100 224027 145414 53634 3816 91808 38471

CV residual 97830 -179813 64036 13409 -7944 -148782 -18208 -65467 15164 -82092
-129379
           170   181   183   187   191   195
Predicted   -6266 25812 10102 10516   85455 50696
cvpred   -6615 73880 28984 29994 217439 132161
PrizeMoney 11421 20064 11309 14098   68613 38043
CV residual 18036 -53816 -17675 -15896 -148826 -94118

Sum of squares = 3.93e+11   Mean square = 1.01e+10   n = 39

Overall (Sum over all 39 folds)
     ms
4.49e+09



> pgatour2006_4$BirdieConversion_SQ<- pgatour2006$BirdieConversion^2
> model_sq2<-lm(PrizeMoney~., data = pgatour2006_4)
> summary(model_sq2)

Call:
lm(formula = PrizeMoney ~ ., data = pgatour2006_4)

Residuals:
    Min   1Q Median   3Q   Max
-132258 -21113   -2549   15657 248898

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)

```
(Intercept)          6611581    1121182    5.90  1.7e-08 ***
DrivingAccuracy        -1094        725   -1.51  0.13293
GIR                  -182224      34982   -5.21  5.0e-07 ***
BirdieConversion      -93306      25373   -3.68  0.00031 ***
SandSaves              1184        648    1.83  0.06919 .
Scrambling             3893       1284    3.03  0.00277 **
GIR_SQ                 1478        269    5.50  1.2e-07 ***
BirdieConversion_SQ    1797        437    4.11  5.9e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43900 on 188 degrees of freedom
Multiple R-squared:  0.546,    Adjusted R-squared:  0.529
F-statistic: 32.3 on 7 and 188 DF,  p-value: <2e-16

> out_2<- cv.lm(data = pgatour2006_4 , form.lm = (PrizeMoney~.),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    0.25 0.61615
GIR                 1 1.54e+11 1.54e+11   80.12 3.4e-16 ***
BirdieConversion      1 1.10e+11 1.10e+11   57.29 1.6e-12 ***
SandSaves             1 3.56e+10 3.56e+10   18.49 2.7e-05 ***
Scrambling            1 2.32e+10 2.32e+10   12.09 0.00063 ***
GIR_SQ                1 7.87e+10 7.87e+10   40.94 1.2e-09 ***
BirdieConversion_SQ   1 3.25e+10 3.25e+10   16.89 5.9e-05 ***
Residuals           188 3.62e+11 1.92e+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
Predicted    52661 60317  82429 55140 -3808  39983  83309  48853  54475 16458  38554
93558
cvpred       50903 54236  78069 57981 13143  42735  78712  39486  56600 27605  30496
82574
PrizeMoney   17516 57273  49640 53610 11989  28658  33471   8734  45752 31371  13262
132327
CV residual -33387  3037 -28429 -4371 -1154 -14077 -45241 -30752 -10848  3766 -17234
49753
```

```
            54     59    61     63    69     81     92    94     96     98    102   107
Predicted    32879  89220 29421  69002  7350   2688  75466 42123  37276  40495 193616
75280
cvpred       28342  83207 32852  64854 21285  17077  57883 37529  34323  47319 149276
65779
PrizeMoney   13865  57092 54477 217748 15840   5265 100398 27673   9149  15964  70421
91406
CV residual -14477 -26115 21625 152894 -5445 -11812  42515 -9856 -25174 -31355 -78855
25627
           123    130    144   152    156   163    165    167   171   176    177    178
Predicted   33714  69567  26822 -7104  50283 50791  41320  60093 31702 27462  33537
413873
cvpred      38000  68318  34489  4963  57253 51134  41971  61162 39590 33690  27670
272630
PrizeMoney  41390  56693  24379 10715  36428 56305  19997  27657 36289 36861   9062
662771
CV residual  3390 -11625 -10110  5752 -20825  5171 -21974 -33505 -3301  3171 -18608
390141
           180   186   196
Predicted     581 73283 29503
cvpred      10491 70039 33306
PrizeMoney  65783 72623 90824
CV residual 55292  2584 57518

Sum of squares = 2.05e+11    Mean square = 5.25e+09    n = 39

fold 2
Observations in test set: 40
             7    12    13    18    20     26    30     34    37     38    44     47
Predicted   44429 37225 36290 26714 26332 101518 34176  59950 57110  46889 25146
102075
cvpred      55132 31407 37826 28887 23260 103947 26656  55103 49690  46116 27840
100869
PrizeMoney  50620 44080 47172 20911 19683  33782 94571  37735 59151  18345 38275
10504
CV residual -4512 12673  9346 -7976 -3577 -70165 67915 -17368  9461 -27771 10435 -90365
            50     52     58    60     70     85    87     89     93    97   109    110
Predicted    32002  78084  58422  9110  21868  42060 63756 64639  72054 -6253 34257
79640
cvpred       31239  63420  42978   728  35324  31176 61633 49101  59579  1322 33708  74681
PrizeMoney   15187 119444 129234 45904   2240  20612 56058 54513  37004  2692 26899
25918
CV residual -16052  56024  86256 45176 -33084 -10564 -5575  5412 -22575  1370 -6809
-48763
```

```
           113    116    117    121    128    134    153   159    166     168    173
Predicted    13127 103154 101330  23336  15550  39951  62174 60925  69213  117350  83535
cvpred       23179  89254  87992  21623  16679  43083  45775 58506  56717  120508  66460
PrizeMoney   12110  83483 176523  11315   5285  26532 119240 69173 114055   15012
105997
CV residual -11069  -5771  88531 -10308 -11394 -16551  73465 10667  57338 -105496  39537
           174    182    185    188    192
Predicted    77233  20447  9101 128284 136005
cvpred       61215  21601 10056 114667 114841
PrizeMoney  150889  11187 84604 160175 170460
CV residual  89674 -10414 74548  45508  55619

Sum of squares = 8.53e+10    Mean square = 2.13e+09    n = 40

fold 3
Observations in test set: 39
             2     9     10     28     33     36     42     53     62     64     65     66
Predicted   122926  5180  42815  83636 27694 15672 23309 50057 29733  38716  -9726
89679
cvpred      103577  -642  36250  82166 21801 15124 23121 38926 25020  34302 -11919  80488
PrizeMoney  262045 86782  23396  37751 51770 50249 14499 73819 43820   5402  10528
54862
CV residual 158468 87424 -12854 -44415 29969 35125 -8622 34893 18800 -28900  22447
-25626
           68     73    74     77    78    104   105    106   108   111    115    132
Predicted   42716  45053 61462  72292 13567  45584  1313 119992 54271 12655  24068
86120
cvpred      22999  33883 54738  74375 14999  35350 -3837 104224 35832  4214  25282  84361
PrizeMoney  39356 103594 57216  36918  7583 117801 30068  58189 37214 42589   3025
42890
CV residual 16357  69711  2478 -37457 -7416  82451 33905 -46035  1382 38375 -22257
-41471
           135   137   138   140   146   148    150    151   157   160   169    172
Predicted   77038   661 31499 23334 92638  8587  54976 -18782 43593 32667 53511 137136
cvpred      67015  1366 28415 16231 74903 -3480  45460 -20465 35783 23445 40597 115997
PrizeMoney  89312 11376 23403 14527 68345 16455 111028   4667 32843 47046 42958
106577
CV residual 22297 10010 -5012 -1704 -6558 19935  65568  25132 -2940 23601  2361  -9420
           175   184   193
Predicted    57510  9370 18224
cvpred       55593 12185 11945
PrizeMoney   15098  6117 12803
CV residual -40495 -6068   858
```

Sum of squares = 6.93e+10 Mean square = 1.78e+09 n = 39

fold 4
Observations in test set: 39
```
                1     5    11     21     35    49     55     56    57     71    72    75
Predicted   82837 27723 41329  97237  92948 73565  46765  50570 53801  74302 -3570 49528
cvpred      75139 25901 39037 100000  91186 67485  51909  51393 44618  69306 -2740 52565
PrizeMoney  60661 16683 29567  79316  38455 65174  26301  22340 43951  38188 13031 82196
CV residual -14478 -9218 -9470 -20684 -52731 -2311 -25608 -29053  -667 -31118 15771 29631
               79    80     83    84    86     88     90    91    95    100   120   122
Predicted   47724 18875  159619 -5706 50579  60987 169172  8708 31630 137226 24141 90451
cvpred      43686 18857  171913 -8281 48992  59311 159532 14720 24568 147309 13589 90958
PrizeMoney  57824 24724   27361 55014 43173  19594 300555  7331 29296  58953 26123 18513
CV residual 14138  5867 -144552 63295 -5819 -39717 141023 -7389  4728 -88356 12534 -72445
              124   125   129    133   136   141   147    149   155   158   162   179
Predicted   36429  35157 81144  -1501 30235 11641 10870  55868 42375 17811 29515 64254
cvpred      32693  34247 78762 -11672 30196 15532  9351  53969 33344 17834 19704 55775
PrizeMoney  22467   7490 78489  25135 37869 38046 14558  19200 51005 19973 20502 89770
CV residual -10226 -26757  -273  36807  7673 22514  5207 -34769 17661  2139   798 33995
              189   190    194
Predicted   40179  9929  44848
cvpred      42118 11051  44540
PrizeMoney  55581 10354  30344
CV residual 13463  -697 -14196
```

Sum of squares = 7.3e+10 Mean square = 1.87e+09 n = 39

fold 5
Observations in test set: 39
```
                3      6      16      22     23     24     27      29     31      32
Predicted    7826  100634   86842  114500  12354  28892  31643   51152  23183   69933
cvpred      25870  309761  273476  355766  55355  98837 107004  164642  85516  225847
PrizeMoney   3635  107294   26129  120927  24814  27224  20322   60073  15668  112443
CV residual -22235 -202467 -247347 -234839 -30541 -71613 -86682 -104569 -69848 -113404
               40     45     46      67    76      82     99    101     103     112
Predicted    54546  151922  11730   90541   572   43751  36885   9963   44902   43083
```

```
cvpred       190371  473305  44701  284516 16132  139483 118591  53529  144511  141874
PrizeMoney    56873   46377  16630   30656 25804   16927  53530   2426   18085   18494
CV residual -133498 -426928 -28071 -253860  9672 -122556 -65061 -51103 -126426 -123380
             114     118     119     126     127   131     139     142     143     145     154
Predicted    28616   56939  -1343   83798 -20438 -4718  14190  219176  57883  37071 -11794
cvpred       98285  184244  16956  266456 -46145 -1577  54377  670890 186447 122341
-26415
PrizeMoney   18721   20188   5777   18838   4444  8272  37100  224027 145414  53634   3816
CV residual -79564 -164056 -11179 -247618  50589  9849 -17277 -446863 -41033 -68707
30231
             161     164   170     181   183     187     191     195
Predicted     63965   62382  -812  20420  1477   6448  130656   45579
cvpred       201921  199121 12920  74999 11449  28063  409418  147068
PrizeMoney    91808   38471 11421  20064 11309  14098   68613   38043
CV residual -110113 -160650 -1499 -54935  -140 -13965 -340805 -109025

Sum of squares = 9.92e+11    Mean square = 2.54e+10    n = 39

Overall (Sum over all 39 folds)
      ms
7.27e+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.

By using backward selection, the initial model and final model would be as below, which is different from the model that we build up in b :

Initial Model:

```
PrizeMoney ~ AveDrivingDistance + DrivingAccuracy + GIR + PuttingAverage +
    BirdieConversion + SandSaves + Scrambling + BounceBack +
    PuttsPerRound + AveDrivingDistance_SQ + DrivingAccuracy_SQ +
    GIR_SQ + PuttingAverage_SQ + BirdieConversion_SQ + SandSaves_SQ +
    Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ
```

Final Model:

```
PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + BounceBack +
    PuttsPerRound + GIR_SQ + BirdieConversion_SQ + SandSaves_SQ +
    Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ
```

We start with the full model with k variables and remove variables one at a time until we reach a threshold for the r-squared. We got adjusted r-squared with 0.546 , which is a little better than the one that we got in question b.

If we evaluated the model by using 5 fold cross validation, the overall average of the mean square is 6.25e+09 , which is smaller than the model that we built in question b .

```
> model_full<- lm(PrizeMoney~. , data = pgatour2006)
> step<- stepAIC(model_full, direction="backward")
Start:  AIC=4205
PrizeMoney ~ AveDrivingDistance + DrivingAccuracy + GIR + PuttingAverage +
    BirdieConversion + SandSaves + Scrambling + BounceBack +
    PuttsPerRound + AveDrivingDistance_SQ + DrivingAccuracy_SQ +
    GIR_SQ + PuttingAverage_SQ + BirdieConversion_SQ + SandSaves_SQ +
    Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ

                      Df Sum of Sq      RSS  AIC
- DrivingAccuracy_SQ     1  8.91e+07 3.36e+11 4203
- DrivingAccuracy        1  2.69e+08 3.36e+11 4203
- AveDrivingDistance     1  6.06e+08 3.36e+11 4204
- SandSaves              1  6.74e+08 3.37e+11 4204
- AveDrivingDistance_SQ  1  6.78e+08 3.37e+11 4204
- SandSaves_SQ           1  1.02e+09 3.37e+11 4204
- PuttingAverage_SQ      1  1.18e+09 3.37e+11 4204
- PuttingAverage         1  1.20e+09 3.37e+11 4204
- Scrambling             1  1.23e+09 3.37e+11 4204
- Scrambling_SQ          1  1.62e+09 3.37e+11 4204
<none>                              3.36e+11 4205
- PuttsPerRound          1  4.17e+09 3.40e+11 4206
- BounceBack             1  4.18e+09 3.40e+11 4206
- PuttsPerRound_SQ       1  4.21e+09 3.40e+11 4206
- BounceBack_SQ          1  4.59e+09 3.40e+11 4206
- BirdieConversion       1  1.87e+10 3.55e+11 4214
- BirdieConversion_SQ    1  2.29e+10 3.59e+11 4216
- GIR                    1  3.66e+10 3.72e+11 4224
- GIR_SQ                 1  4.17e+10 3.78e+11 4226

Step:  AIC=4203
PrizeMoney ~ AveDrivingDistance + DrivingAccuracy + GIR + PuttingAverage +
    BirdieConversion + SandSaves + Scrambling + BounceBack +
    PuttsPerRound + AveDrivingDistance_SQ + GIR_SQ + PuttingAverage_SQ +
    BirdieConversion_SQ + SandSaves_SQ + Scrambling_SQ + BounceBack_SQ +
    PuttsPerRound_SQ
```

```
                      Df Sum of Sq      RSS  AIC
- AveDrivingDistance     1  5.21e+08 3.36e+11 4202
- AveDrivingDistance_SQ  1  5.91e+08 3.37e+11 4202
- SandSaves              1  6.49e+08 3.37e+11 4202
- SandSaves_SQ           1  9.90e+08 3.37e+11 4202
- PuttingAverage_SQ      1  1.13e+09 3.37e+11 4202
- PuttingAverage         1  1.15e+09 3.37e+11 4202
- Scrambling             1  1.35e+09 3.37e+11 4202
- Scrambling_SQ          1  1.76e+09 3.38e+11 4202
<none>                               3.36e+11 4203
- BounceBack             1  4.11e+09 3.40e+11 4204
- PuttsPerRound          1  4.42e+09 3.40e+11 4204
- PuttsPerRound_SQ       1  4.46e+09 3.40e+11 4204
- BounceBack_SQ          1  4.51e+09 3.40e+11 4204
- DrivingAccuracy        1  6.45e+09 3.42e+11 4205
- BirdieConversion       1  1.86e+10 3.55e+11 4212
- BirdieConversion_SQ    1  2.28e+10 3.59e+11 4214
- GIR                    1  3.95e+10 3.75e+11 4223
- GIR_SQ                 1  4.47e+10 3.81e+11 4226

Step:  AIC=4202
PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion +
    SandSaves + Scrambling + BounceBack + PuttsPerRound + AveDrivingDistance_SQ +
    GIR_SQ + PuttingAverage_SQ + BirdieConversion_SQ + SandSaves_SQ +
    Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ

                      Df Sum of Sq      RSS  AIC
- SandSaves              1  6.94e+08 3.37e+11 4200
- PuttingAverage_SQ      1  9.08e+08 3.37e+11 4200
- PuttingAverage         1  9.25e+08 3.37e+11 4200
- SandSaves_SQ           1  1.06e+09 3.38e+11 4200
- Scrambling             1  1.09e+09 3.38e+11 4200
- Scrambling_SQ          1  1.47e+09 3.38e+11 4201
- AveDrivingDistance_SQ  1  1.48e+09 3.38e+11 4201
<none>                               3.36e+11 4202
- BounceBack             1  3.85e+09 3.40e+11 4202
- BounceBack_SQ          1  4.25e+09 3.41e+11 4202
- PuttsPerRound          1  4.56e+09 3.41e+11 4202
- PuttsPerRound_SQ       1  4.59e+09 3.41e+11 4202
- DrivingAccuracy        1  6.58e+09 3.43e+11 4203
- BirdieConversion       1  1.82e+10 3.55e+11 4210
- BirdieConversion_SQ    1  2.24e+10 3.59e+11 4212
- GIR                    1  4.00e+10 3.77e+11 4222
```

```
- GIR_SQ                1  4.53e+10 3.82e+11 4224

Step:  AIC=4200
PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion +
    Scrambling + BounceBack + PuttsPerRound + AveDrivingDistance_SQ +
    GIR_SQ + PuttingAverage_SQ + BirdieConversion_SQ + SandSaves_SQ +
    Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ

                      Df Sum of Sq      RSS  AIC
- PuttingAverage_SQ      1  8.86e+08 3.38e+11 4199
- PuttingAverage         1  9.04e+08 3.38e+11 4199
- AveDrivingDistance_SQ  1  1.54e+09 3.39e+11 4199
- Scrambling             1  1.80e+09 3.39e+11 4199
- Scrambling_SQ          1  2.29e+09 3.39e+11 4199
<none>                              3.37e+11 4200
- BounceBack             1  3.96e+09 3.41e+11 4200
- PuttsPerRound          1  3.98e+09 3.41e+11 4200
- PuttsPerRound_SQ       1  4.02e+09 3.41e+11 4200
- BounceBack_SQ          1  4.37e+09 3.42e+11 4201
- SandSaves_SQ           1  5.18e+09 3.42e+11 4201
- DrivingAccuracy        1  6.52e+09 3.44e+11 4202
- BirdieConversion       1  1.80e+10 3.55e+11 4208
- BirdieConversion_SQ    1  2.22e+10 3.59e+11 4211
- GIR                    1  3.96e+10 3.77e+11 4220
- GIR_SQ                 1  4.49e+10 3.82e+11 4223

Step:  AIC=4199
PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion +
    Scrambling + BounceBack + PuttsPerRound + AveDrivingDistance_SQ +
    GIR_SQ + BirdieConversion_SQ + SandSaves_SQ + Scrambling_SQ +
    BounceBack_SQ + PuttsPerRound_SQ

                      Df Sum of Sq      RSS  AIC
- PuttingAverage         1  1.97e+08 3.38e+11 4197
- AveDrivingDistance_SQ  1  1.41e+09 3.39e+11 4197
- Scrambling             1  1.89e+09 3.40e+11 4198
- Scrambling_SQ          1  2.37e+09 3.40e+11 4198
<none>                              3.38e+11 4199
- BounceBack             1  3.91e+09 3.42e+11 4199
- BounceBack_SQ          1  4.32e+09 3.42e+11 4199
- SandSaves_SQ           1  4.97e+09 3.43e+11 4199
- DrivingAccuracy        1  6.03e+09 3.44e+11 4200
- PuttsPerRound          1  1.17e+10 3.50e+11 4203
- PuttsPerRound_SQ       1  1.18e+10 3.50e+11 4203
```

```
- BirdieConversion       1  2.19e+10 3.60e+11 4209
- BirdieConversion_SQ    1  2.82e+10 3.66e+11 4212
- GIR                    1  4.85e+10 3.87e+11 4223
- GIR_SQ                 1  5.51e+10 3.93e+11 4226

Step:  AIC=4197
PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + Scrambling +
    BounceBack + PuttsPerRound + AveDrivingDistance_SQ + GIR_SQ +
    BirdieConversion_SQ + SandSaves_SQ + Scrambling_SQ + BounceBack_SQ +
    PuttsPerRound_SQ

                       Df Sum of Sq      RSS  AIC
- AveDrivingDistance_SQ  1  1.22e+09 3.39e+11 4195
- Scrambling             1  1.84e+09 3.40e+11 4196
- Scrambling_SQ          1  2.39e+09 3.41e+11 4196
<none>                              3.38e+11 4197
- BounceBack             1  4.02e+09 3.42e+11 4197
- BounceBack_SQ          1  4.48e+09 3.43e+11 4197
- SandSaves_SQ           1  5.11e+09 3.43e+11 4198
- DrivingAccuracy        1  5.98e+09 3.44e+11 4198
- PuttsPerRound          1  1.18e+10 3.50e+11 4201
- PuttsPerRound_SQ       1  1.19e+10 3.50e+11 4201
- BirdieConversion       1  2.22e+10 3.60e+11 4207
- BirdieConversion_SQ    1  2.81e+10 3.66e+11 4210
- GIR                    1  4.86e+10 3.87e+11 4221
- GIR_SQ                 1  5.49e+10 3.93e+11 4224

Step:  AIC=4195
PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + Scrambling +
    BounceBack + PuttsPerRound + GIR_SQ + BirdieConversion_SQ +
    SandSaves_SQ + Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ

                    Df Sum of Sq      RSS  AIC
- Scrambling          1  1.83e+09 3.41e+11 4194
- Scrambling_SQ       1  2.37e+09 3.42e+11 4195
<none>                           3.39e+11 4195
- BounceBack          1  4.03e+09 3.43e+11 4196
- BounceBack_SQ       1  4.48e+09 3.44e+11 4196
- DrivingAccuracy     1  5.07e+09 3.45e+11 4196
- SandSaves_SQ        1  6.13e+09 3.46e+11 4197
- PuttsPerRound       1  1.19e+10 3.51e+11 4200
- PuttsPerRound_SQ    1  1.20e+10 3.51e+11 4200
- BirdieConversion    1  2.19e+10 3.61e+11 4206
- BirdieConversion_SQ 1  2.74e+10 3.67e+11 4209
```

```
- GIR                1  5.07e+10 3.90e+11 4221
- GIR_SQ             1  5.69e+10 3.96e+11 4224

Step:  AIC=4194
PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + BounceBack +
    PuttsPerRound + GIR_SQ + BirdieConversion_SQ + SandSaves_SQ +
    Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ

                   Df Sum of Sq      RSS  AIC
<none>                           3.41e+11 4194
- BounceBack           1  4.80e+09 3.46e+11 4195
- DrivingAccuracy      1  4.93e+09 3.46e+11 4195
- BounceBack_SQ        1  5.24e+09 3.47e+11 4195
- SandSaves_SQ         1  6.23e+09 3.48e+11 4196
- Scrambling_SQ        1  8.60e+09 3.50e+11 4197
- PuttsPerRound        1  1.06e+10 3.52e+11 4198
- PuttsPerRound_SQ     1  1.06e+10 3.52e+11 4198
- BirdieConversion     1  2.19e+10 3.63e+11 4205
- BirdieConversion_SQ  1  2.76e+10 3.69e+11 4208
- GIR                  1  5.26e+10 3.94e+11 4221
- GIR_SQ               1  5.88e+10 4.00e+11 4224
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
PrizeMoney ~ AveDrivingDistance + DrivingAccuracy + GIR + PuttingAverage +
    BirdieConversion + SandSaves + Scrambling + BounceBack +
    PuttsPerRound + AveDrivingDistance_SQ + DrivingAccuracy_SQ +
    GIR_SQ + PuttingAverage_SQ + BirdieConversion_SQ + SandSaves_SQ +
    Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ

Final Model:
PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion + BounceBack +
    PuttsPerRound + GIR_SQ + BirdieConversion_SQ + SandSaves_SQ +
    Scrambling_SQ + BounceBack_SQ + PuttsPerRound_SQ


                  Step Df Deviance Resid. Df Resid. Dev  AIC
1                                      177   3.36e+11 4205
2    - DrivingAccuracy_SQ  1 8.91e+07       178   3.36e+11 4203
3    - AveDrivingDistance  1 5.21e+08       179   3.36e+11 4202
4            - SandSaves   1 6.94e+08       180   3.37e+11 4200
5     - PuttingAverage_SQ  1 8.86e+08       181   3.38e+11 4199
```

6	- PuttingAverage	1	1.97e+08	182	3.38e+11 4197
7	- AveDrivingDistance_SQ	1	1.22e+09	183	3.39e+11 4195
8	- Scrambling	1	1.83e+09	184	3.41e+11 4194

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.

Produce a model using backward selection, we would have a better adjusted r-squared and got a better average mean squared according to the 5-fold CV, so I would say the backward selection is better in this case.
However, feature selection is just a tool, human beings have to choose the features that go into the model, if we are in a domain where speed is the most important, then we should use backward elimination to build up the model.