Rhea Qianqi Rao
Regression and Multivariate Data Analysis
Professor Simonoff
Dec 19th, 2015

What makes some movies more likely to win awards?

Being a big fan of movies, I watch a movie almost every week, and I also keep up to date with the Film Awards like the Oscar. However, it is almost always a disappointment watching the academy ceremony, since it is hardly the case that my favorite movie wins the award. In this assignment, I want to explore the reason why certain movies win awards while other movies that are of similar caliber and have fair shots at winning, do not. I manually collected a data of 58 movies, which are movies that show up on the lists of most popular movies in 2014 on different websites such as Yahoo, Rotten Tomatoes, and Metacritic. These movies are all released during Jan 2014 to Jan 2015, and they are all made/releases in United States. These are the movies that have gathered a lot of momentum among the audience – but only some of them won awards of some kind. Then I went on Metacritics, where there is a scorecard of all the film awards from over 300 organizations that are given out in 2014. Within my list, a movie is categorized as Award-winning if it was affiliated with an award in any way – be it Best Picture, or Best Actor, etc.

After dividing the movies into ones that won awards of some kind, and ones that did not, three metrics are chosen as potential predictors of award winning. I originally wanted to include the budget of the movie as an indicator as well. However, such information was a bit hard to obtain, so I just went with these three predictors.

- Release Days (**Days**): this refers to the days a certain movie has been released since Jan 1st, 2014. From my past experiences of watching Oscar ceremonies, I suspected that movies that are released earlier in the year (Jan-Mar) appear to have a lesser chance of winning the award, so I list the release days as one of the predictor. I expect movies that won awards to have a higher number of release days.

- Rotten Tomatoes Scores (**RT Score**): this refers to the score a certain movie receives on Rotten Tomatoes (shown as a number of 100). As these movies are all popular movies that have gather large momentum, they all have relative high scores with an average of 88 out of 100. However, we still expect to see that the award winning movies to be the more popular ones, and hence having higher RT Scores.

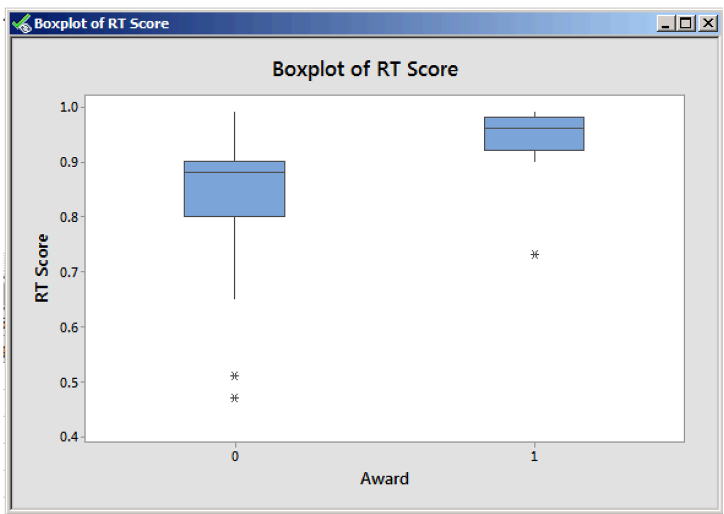
- Length Of The Movie (**Length**): this refers to how long the movie lasts in minutes, since the award movie judges have to sit through different movies in one setting, it might be that a movie that is longer in length would be less favorable I judges' eyes. In general I expect to see the award-winning movies to have a short length of time.

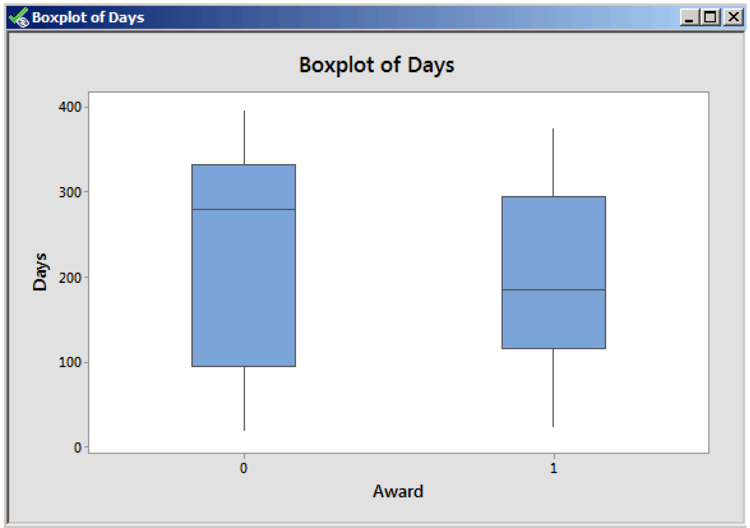
Here are all the data:

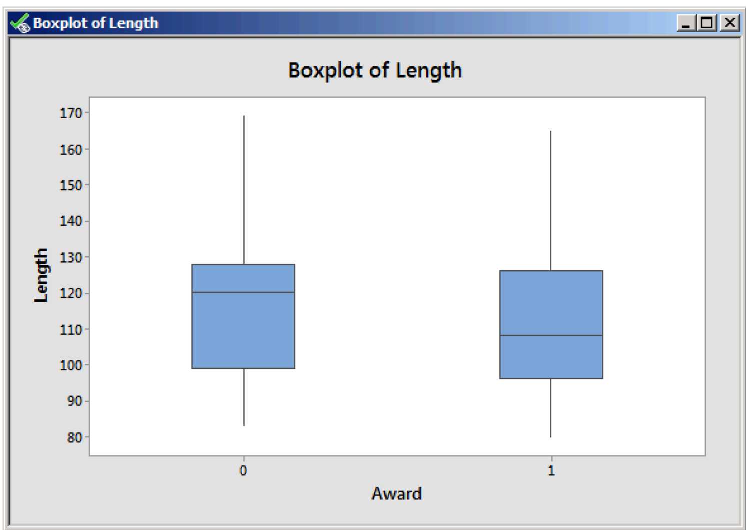
Movies	Award	RT Score	Days	Length
BOYHOOD	1	0.98	162	165
MR. TURNER	1	0.98	353	149
THE BABADOOK	1	0.98	332	94
THE LEGO MOVIE	1	0.96	38	101
LIFE ITSELF	1	0.97	185	120
WHIPLASH	1	0.94	283	106
NIGHTCRAWLER	1	0.95	304	117
GLORIA	1	0.99	24	108
LEVIATHAN	1	0.99	148	130
TWO DAYS, ONE NIGHT	1	0.97	358	95
SNOWPIERCER	1	0.95	294	126
CITIZENFOUR	1	0.98	294	114
STARRED UP	1	0.99	239	106
BIRDMAN	1	0.92	288	119
THE GRAND BUDAPEST HOTEL	1	0.92	66	99
JODOROWSKY'S DUNE	1	0.98	80	90
X-MEN DAY OF THE PAST	1	0.91	143	131
GUARDIANSOF THE GALAXY	1	0.91	213	121
WE ARE THE BEST	1	0.97	150	102
IDA	1	0.96	122	80
LIVE DIES REPEAT	1	0.9	157	113
BLUE RUIN	1	0.96	115	90
DAWN OF THE PLANET OF THE APES	1	0.9	192	130
THE MISSING PICTURE	1	0.99	78	96
SONG OF THE SEA	1	0.99	353	93
THE LUNCHBOX	1	0.96	59	104
INHERENT VICE	1	0.73	374	148
GONE GIRL	0	0.9	276	145
THE IMMITATION GAME	0	0.9	332	114
WILD	0	0.9	337	115
BIG HERO 6	0	0.9	311	93
THE FAULTS IN OUR STARS	0	0.8	127	125
22 JUMP STREET	0	0.85	164	110
THE HUNGER GAMES: MOCKING JAY	0	0.65	325	125
THE THEORY OF EVERYTHING	0	0.8	311	123
FOXCATCHER	0	0.88	318	130
BEGIN AGAIN	0	0.83	178	101
BLUE JASMINE	0	0.91	21	98
AMERICAN SNIPER	0	0.72	381	134
INTERSTELLAR	0	0.71	311	169
INTO THE WOODS	0	0.71	83	125
THE JUDGE	0	0.47	27	142
SELMA	0	0.99	374	127
STILL ALICE	0	0.88	381	99

UNBROKEN	0	0.51	83	137
THE OVERNIGHTERS	0	0.98	283	90
FORCE MAJEURE	0	0.92	41	120
DEAR WHITE PEOPLE	0	0.91	290	108
A MOST VIOLENT YEAR	0	0.89	395	125
COHERENCE	0	0.88	20	89
TOP FIVE	0	0.86	346	101
OBVIOUS CHILD	0	0.89	280	83
BEYOND THE LIGHTS	0	0.81	55	102
THE SKELETON TWINS	0	0.87	350	93
ONLY LOVERS LEFT ALIVE	0	0.86	231	122
CAPTAIN AMERICA: THE WINTER SOLDIER	0	0.89	252	136
UNDER THE SKIN	0	0.85	94	128
THE GUEST	0	0.9	260	97

Since constant variance isn't a requisite in the logistic regression model, I didn't take log of any kind for the data. Looking at the data, we will first construct side-by-side boxplots to see if there is clear separation between the two groups on the variables. Note that this does not take into account the variables having joint effects:







The RT Score variable shows clear separation between Award-winning and Non-award-winning movies as expected. The Days variable shows less predictive power. We were expecting the award winning movies to be released later in date and thus having a bigger number in days, but it shows in the graph that the award-winning ones actually have a lower Days. Award-winning movies also have a generally lower Length in time, although other aspects of the graph – tails and size of the boxes are pretty similar. None of the graphs are particularly skewed, the boxplots of length might be slightly right skewed, but it is only slightly so, and the notion of non-constant variance is not relevant, so the data is the good the way it is.

Now let's try to fit a logistic regression model to fit the data:

Binary Logistic Regression: Award versus RT Score, Release Date, Length Of Movie

```
Method
```

Link function	Logit
Rows used	58

Response Information

Variable	Value	Count	
Award	1	27	(Event)
	0	31	
	Total	58	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	3	31.902	10.6340	31.90	0.000
RT Score	1	30.455	30.4553	30.46	0.000
Release Date	1	1.871	1.8712	1.87	0.171
Length Of Movie	1	2.136	2.1355	2.14	0.144
Error	54	48.227	0.8931		
Total	57	80.129			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
39.81%	36.07%	56.23

Coefficients

Term	Coef	SE Coef	VIF
Constant	-31.45	9.21	
RT Score	0.3146	0.0877	1.15
Release Date	-0.00445	0.00334	1.06
Length Of Movie	0.0323	0.0234	1.19

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
RT Score	1.3697	(1.1533, 1.6266)
Release Date	0.9956	(0.9891, 1.0021)
Length Of Movie	1.0329	(0.9866, 1.0813)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -31.45 +†0.3146†RT†Score -†0.00445†Release†Date +†0.0323†Length†Of†Movie

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	54	48.23	0.696
Pearson	54	246.53	0.000
Hosmer-Lemeshow	8	11.65	0.167

Observed and Expected Frequencies for Hosmer-Lemeshow Test

Group	Event Probability Range	Award = 1 Observed	Award = 1 Expected	Award = 0 Observed	Award = 0 Expected
1	(0.000, 0.004)	0	0.0	5	5.0
2	(0.004, 0.057)	1	0.2	5	5.8
3	(0.057, 0.180)	0	0.7	6	5.3
4	(0.180, 0.273)	0	1.5	6	4.5
5	(0.273, 0.514)	2	2.3	4	3.7
6	(0.514, 0.579)	3	2.8	2	2.2
7	(0.579, 0.719)	6	4.0	0	2.0
8	(0.719, 0.844)	4	4.6	2	1.4
9	(0.844, 0.893)	6	5.2	0	0.8
10	(0.893, 0.982)	5	5.6	1	0.4

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	755	90.2	Somers' D	0.81
Discordant	79	9.4	Goodman-Kruskal Gamma	0.81
Ties	3	0.4	Kendall's Tau-a	0.41
Total	837	100.0		

Association is between the response variable and predicted probabilities

Fits and Diagnostics for Unusual Observations

Obs	Observed Probability	Fit	Resid	Std Resid	
27	1.0000	0.0046	3.2780	3.30	R
43	0.0000	0.8940	-2.1187	-2.21	R

R Large residual

Looking at the coefficients section, we see that when holding everything else constant, 1 unit increase in RT Score (which is 1 out of 100 full score), is associated with 0.31% higher odds of winning an award. Similarly, if a movie is released a day later in the year, this is associated with 0.44% lower odds of winning an award of any kind. The odds ratio is the exponential factor of the coefficient.

Checking to see if the data is significant, we look at the Deviance Table, which shows us the likelihood ratio test. We can see that the regression has a Chi-square of 31.9, and a p-value of 0 out to 3 digits. So we know the overall regression is statistically significant. There is no physical interpretation for R-square. For individual variables, we realize that the RT Score is statistically significant with a p-value of 0 out to 3 digits, when Days and Length coefficients are not that significant. We'll look deeper into this later. Also note that collinearity is not much of a problem here, with VIFs all close to 1.

Looking at the Goodness-of-Fit Tests, Pearson, with a P-value of 0.00, has strong evidence to reject the null hypothesis that the model fits. We see a huge difference between Deviance and Pearson, but we don't care since there is only one replication for each movie, nj =1, these two statistic are thus meaningless. So we look to

Hosmer-Lemeshow statistic, even though the evidence is somewhat weak, we still decided to not reject the null, and accept the model. Somers' D is a strong 0.81.

Since the predictor Length doesn't seem to have as big an impact on our response variable, we want to try to see if taking it out would affect our model at all, so we run a stepwise regression.

Binary Logistic Regression: Award versus Length, RT Score, Days

```
Method

Link function  Logit
Rows used      58


Stepwise Selection of Terms

α to enter = 0.15, α to remove = 0.15


Response Information

Variable  Value  Count
Award     1         27  (Event)
          0         31
          Total     58


Deviance Table

Source      DF  Adj Dev  Adj Mean  Chi-Square  P-Value
Regression   1    25.72   25.7249       25.72    0.000
  RT Score   1    25.72   25.7249       25.72    0.000
Error       56    54.40    0.9715
Total       57    80.13


Model Summary

Deviance   Deviance
    R-Sq  R-Sq(adj)    AIC
  32.10%     30.86%  58.40


Coefficients

Term        Coef  SE Coef   VIF
Constant  -23.45     6.77
RT Score  0.2554   0.0733  1.00


Odds Ratios for Continuous Predictors

          Odds Ratio       95% CI
RT Score      1.2910  (1.1183, 1.4904)


Regression Equation
```

```
P(1)  =  exp(Y')/(1 + exp(Y'))


Y' = -23.45 +†0.2554†RT†Score


Goodness-of-Fit Tests

Test             DF  Chi-Square  P-Value
Deviance         56       54.40    0.535
Pearson          56      159.92    0.000
Hosmer-Lemeshow   8       12.24    0.141


Measures of Association

Pairs       Number  Percent  Summary Measures       Value
Concordant     722     86.3  Somers' D               0.76
Discordant      86     10.3  Goodman-Kruskal Gamma   0.79
Ties            29      3.5  Kendall's Tau-a         0.38
Total          837    100.0

Association is between the response variable and predicted probabilities
```

The stepwise model only fits RT Score as the only predictor in the model. We notice that p-value for regression and RT Score are, as before, strongly statistically significant, and AIC went down from 56.23 to 58.40. However, we also notice that Hosmer-Lemeshow statistic went down to 0.141 from 0.167, and Somers' D went down to 0.76 from 0.81 – it doesn't separate as well as the three predictor model did. Given that I only have three predictors to start off, I don't think it is a good idea, nor it is worth sacrificing the separation, to have a single predictor model.

Although the best subset regression wouldn't be appropriate here, I still wonder if maybe having only two predictors would be a better model. So I run a best subset regression just to see. What I got is this:

Best Subsets Regression: Award versus Length, RT Score, Days

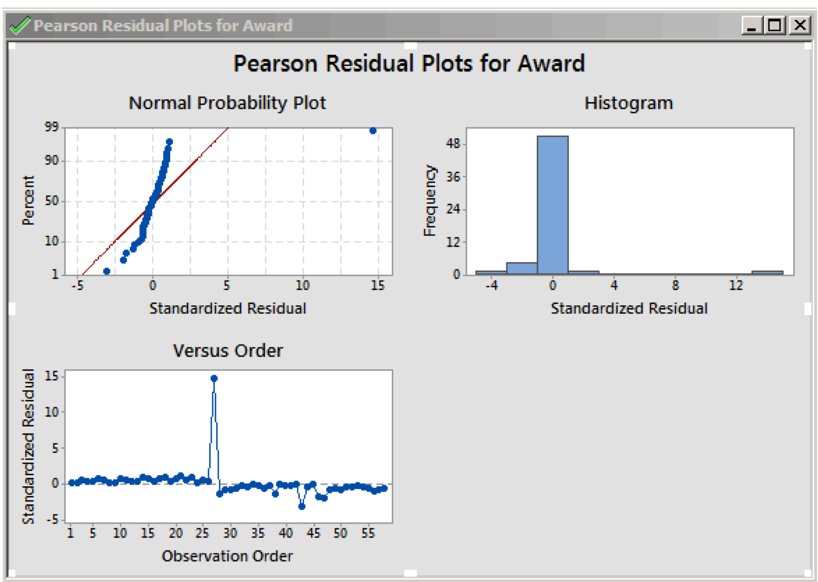
```
Response is Award

                                                       R
                                                       T
                                                     L
                                                     e S
                                                     n c D
                                                     g o a
             R-Sq    R-Sq  Mallows                   t r y
Vars  R-Sq  (adj)  (pred)       Cp        S  h e s
   1  26.7   25.4    20.7      5.4  0.43461    X
   1   2.0    0.3     0.0     25.4  0.50251      X
   2  30.0   27.5    21.7      4.7  0.42852    X X
   2  28.7   26.1    20.3      5.7  0.43241  X X
   3  33.3   29.6    23.2      4.0  0.42213  X X X
```

If this table is to be trusted, we could see that the three-predictor model is the best choice here - it has the highest R-square, lowest Cp, and a leveled out S. So we decided to stay with the three-predictor model.

Now that we decided to stick with the three-predictor model, we want to adjust for possible unusual observations, and here are the regression diagnostics:

Movies	SPEARRES1	HI1	COOK1
BOYHOOD	0.1394	0.047235	0.000241
MR. TURNER	0.2828	0.091849	0.002022
THE BABADOOK	0.6634	0.110499	0.01367
THE LEGO MOVIE	0.4131	0.072511	0.003335
LIFE ITSELF	0.3548	0.044464	0.001464
WHIPLASH	0.8887	0.048146	0.009986
NIGHTCRAWLER	0.6692	0.057025	0.00677
GLORIA	0.2195	0.042419	0.000534
LEVIATHAN	0.2022	0.037929	0.000403
TWO DAYS, ONE NIGHT	0.8186	0.130067	0.025046
SNOWPIERCER	0.569	0.067619	0.005871
CITIZENFOUR	0.428	0.054842	0.002658
STARRED UP	0.3664	0.045291	0.001592
BIRDMAN	0.9976	0.048289	0.012624
THE GRAND BUDAPEST HOTEL	0.8638	0.097708	0.020198
JODOROWSKY'S DUNE	0.3953	0.071215	0.002996
X-MEN DAY OF THE PAST	0.7145	0.095747	0.013513
GUARDIANSOF THE GALAXY	0.9548	0.044743	0.010676
WE ARE THE BEST	0.4395	0.046682	0.002365
IDA	0.7235	0.133513	0.020165
LIVE DIES REPEAT	1.1251	0.048681	0.016194
BLUE RUIN	0.5886	0.081582	0.007694
DAWN OF THE PLANET OF THE APES	0.94	0.080938	0.019456
THE MISSING PICTURE	0.302	0.051645	0.001242
SONG OF THE SEA	0.6089	0.125593	0.013312
THE LUNCHBOX	0.4104	0.063658	0.002863
INHERENT VICE	14.7459	0.013793	0.760281
GONE GIRL	-1.2824	0.16333	0.080261
THE IMMITATION GAME	-0.6477	0.061697	0.006896
WILD	-0.6517	0.06395	0.007255
BIG HERO 6	-0.4887	0.082269	0.005352
THE FAULTS IN OUR STARS	-0.2533	0.062049	0.001061
22 JUMP STREET	-0.4015	0.059952	0.00257
THE HUNGER GAMES: MOCKING JAY	-0.0149	0.001261	0
THE THEORY OF EVERYTHING	-0.1602	0.030246	0.0002
FOXCATCHER	-0.6391	0.082925	0.009234
BEGIN AGAIN	-0.244	0.046846	0.000731
BLUE JASMINE	-1.2235	0.143025	0.062461

AMERICAN SNIPER	-0.046	0.006913	0.000004
INTERSTELLAR	-0.0815	0.023636	0.00004
INTO THE WOODS	-0.0661	0.013887	0.000015
THE JUDGE	-0.0022	0.000071	0
SELMA	-3.0316	0.082298	0.206052
STILL ALICE	-0.3342	0.070089	0.002105
UNBROKEN	-0.0034	0.000138	0
THE OVERNIGHTERS	-1.7594	0.09826	0.084324
FORCE MAJEURE	-1.9137	0.105988	0.108538
DEAR WHITE PEOPLE	-0.7501	0.048515	0.007172
A MOST VIOLENT YEAR	-0.5863	0.098648	0.009405
COHERENCE	-0.6727	0.171683	0.023451
TOP FIVE	-0.2697	0.051205	0.000982
OBVIOUS CHILD	-0.3822	0.09002	0.003613
BEYOND THE LIGHTS	-0.2404	0.065599	0.001014
THE SKELETON TWINS	-0.2763	0.060443	0.001227
ONLY LOVERS LEFT ALIVE	-0.4911	0.058428	0.003741
CAPTAIN AMERICA: THE WINTER SOLDIER	-0.9665	0.105419	0.027518
UNDER THE SKIN	-0.6536	0.133207	0.016414
THE GUEST	-0.5784	0.064316	0.005748



Looking at the three-in-one plot, the plots don't look normally distributed, but that's okay. We notice an obvious unusual observation that is 14 standard deviations away from where it was expected to be. That corresponds to row 27, which is the movie < Inherent Vice>, the residual is over 14, cook distance for this observation is way bigger than the other ones, this is obviously an influential point. <Inherent Vice> has one of the movie that have the lowest RT Scores on the list, and it also has the longest length of time, which according to our model, are not indicative of award winning, yet it did. If we checked the website to see what award it won, it was not surprising to find that it only won one award (when other award-winning movies tend to win more than one), and it was the Best Actor award, which arguably is not that indicative of the movie's quality. So it might be the case that this movie would normally not be that popular, it just happens to have a great actor.

For leverage points, the guideline is 2.5*(3+1)/58 = 0.172. There is one point that is very close to this number, and that's the movie <Coherence>, it is a movie that has average rating, released very early, has a short length, and didn't end up winning any awards. It is almost a leverage point, but it is in general not that big of a violation to the trend of our data, so I decided to not do anything about this point.

So I went back, took <Inherent Vice> out, we get the result:

Binary Logistic Regression: Award versus RT Score, Days, Length

```
* WARNING * When the data are in the Response/Frequency format, the Residuals
versus fits
            plot is unavailable.
```

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	58

Response Information

Variable	Value	Count	
Award	1	26	(Event)
	0	32	
	Total	58	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	3	46.238	15.4128	46.24	0.000
RT Score	1	43.276	43.2756	43.28	0.000
Days	1	3.472	3.4719	3.47	0.062
Length	1	1.364	1.3641	1.36	0.243
Error	54	33.545	0.6212		
Total	57	79.783			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
57.95%	54.19%	41.54

Coefficients

Term	Coef	SE Coef	VIF
Constant	-50.4	14.7	
RT Score	0.525	0.143	1.31
Days	-0.00768	0.00444	1.18
Length	0.0331	0.0298	1.28

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
RT Score	1.6907	(1.2768, 2.2386)
Days	0.9924	(0.9838, 1.0010)
Length	1.0337	(0.9750, 1.0959)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -50.4 +†0.525†RT†Score -†0.00768†Days +†0.0331†Length

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	53	33.54	0.987
Pearson	53	43.46	0.847
Hosmer-Lemeshow	8	5.40	0.714

Observed and Expected Frequencies for Hosmer-Lemeshow Test

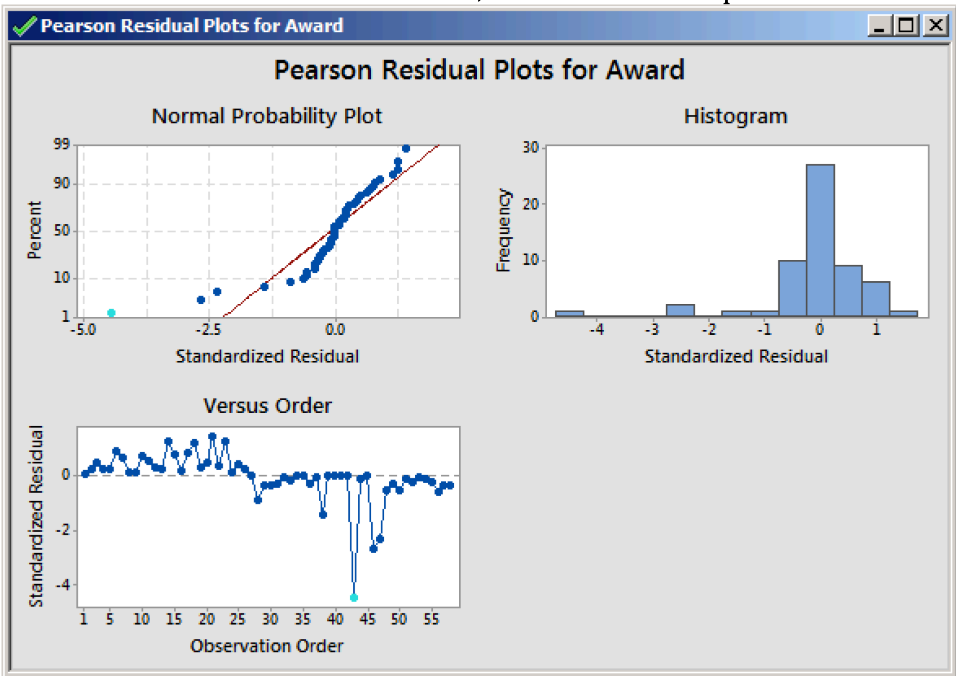
	Event Probability	Award = 1		Award = 0	
Group	Range	Observed	Expected	Observed	Expected
1	(0.000, 0.000)	0	0.0	5	5.0
2	(0.000, 0.008)	0	0.0	6	6.0
3	(0.008, 0.048)	0	0.2	6	5.8
4	(0.048, 0.127)	0	0.6	6	5.4
5	(0.127, 0.386)	1	1.5	5	4.5
6	(0.386, 0.613)	4	2.5	1	2.5
7	(0.613, 0.826)	5	4.4	1	1.6
8	(0.826, 0.929)	5	5.2	1	0.8
9	(0.929, 0.963)	5	5.7	1	0.3
10	(0.963, 0.995)	6	5.9	0	0.1

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	788	94.7	Somers' D	0.89
Discordant	44	5.3	Goodman-Kruskal Gamma	0.89
Ties	0	0.0	Kendall's Tau-a	0.45
Total	832	100.0		

Association is between the response variable and predicted probabilities

After taking out the point, we notice that regression and RT Scores still shows very strong evidence of statistical significance. While Days still doesn't make the 0.05 cut, it is more statistically significant than before. Looking at the Hosmer-Lemeshow statistic, it went from to 0.167 to 0.714, there is stronger evidence to not reject the null. AIC went down from 56 to 41, Somers'D went up to 0.89 as well.



There is still an indication of some usual observations after taking <Inherent Vice> out, but at this point the fit has been greatly improved, and omitting further points might not help us that much.

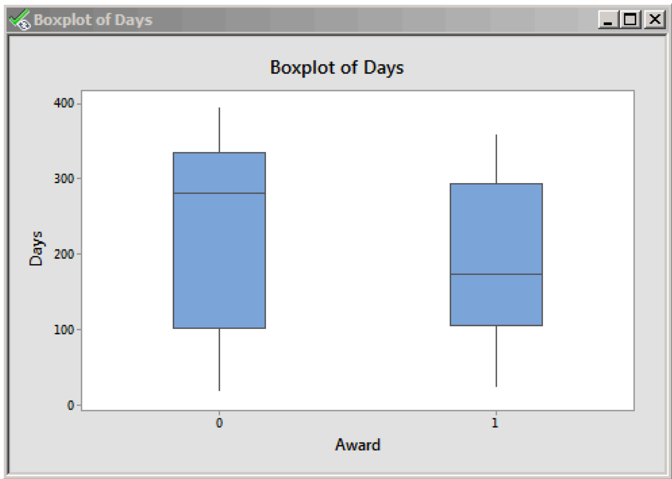
Now that we took out the influential point that seemed to have impacted our model quite a bit, I decided to run best subset again to see if the three predictor model is still the best fit, and we got the following result:

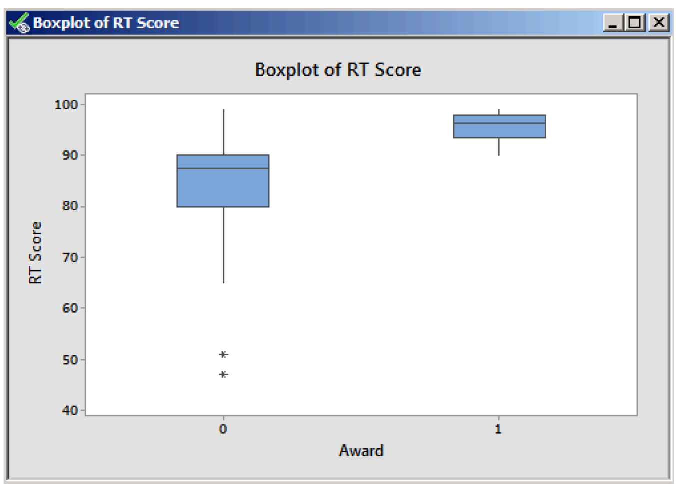
Best Subsets Regression: Award versus Length, RT Score, Days

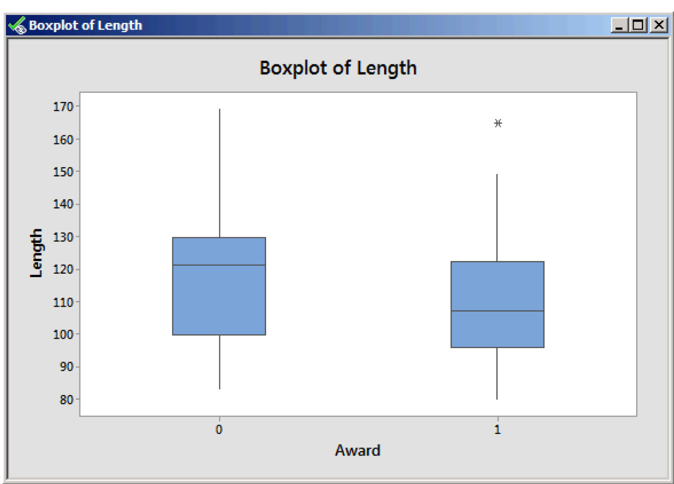
```
Response is Award

                                                    R
                                                    T
                                                  L
                                                  e S
                                                  n c D
                                                  g o a
                R-Sq     R-Sq   Mallows           t r y
Vars  R-Sq    (adj)   (pred)        Cp        S   h e s
   1  32.3     31.1     25.7       6.9  0.41644     X
   1   3.5      1.8      0.0      32.8  0.49706       X
   2  37.7     35.4     30.1       4.0  0.40305     X X
   2  33.3     30.9     24.6       8.0  0.41699   X X
   3  40.0     36.6     30.6       4.0  0.39935   X X X
```

As we can see, with a highest R-sq and lowest Cp, the three-predictor model is still the best fit. Although the two-predictor model with RT Score and Days is arguably nearly as good. In the end, I decided that it is worth it to include one more predictor.

Just out of curiosity, I made the box plots again after taking the influential point out.







The trend we observed earlier in our previous box plots got stronger here.
Finally, now that we have decided on the model, we want to run a classification matrix. I also used the original data with outlier included. The cut off of 0.5 seems reasonable here so we will just use that.

Tabulated Statistics: Award, Predict

```
Rows: Award   Columns: Predict

            0      1     All

0          28      4      32
        48.28   6.90   55.17

1           4     22      26
         6.90  37.93   44.83

All        32     26      58
        55.17  44.83  100.00

Cell Contents:      Count
                    % of Total
```

86.21% of the firms were correctly classified, higher than

$$C_{pro} = (1.25)[(0.5517)(0.5517) + (0.4483)(0.4483)] = 63.17\%$$

We are using the same data twice, but 86.21% is still fairly high. Cmax = 55%, reinforcing the strength of the model. These are all suggesting that our model is better than random chances, in terms of predicting whether a movie would win an award or not.

Conclusion:
When considering the reasons why some movies win awards when movies of the same caliber all have a fair shot at winning, we realized that the Rotten Tomatoes Score of a movie is useful at predicting/classifying the award-winning ones from the others. Although the other two predictors - Days and Length, are not as statistically significant, they still help classify the two groups. We learned that the award-winning movies have a lower mean than the other group, meaning that they are usually released earlier in the year and they in general have a short length of movie time. For the movies that are of certain caliber, the three-predictor binary logistic regression model I have here does reasonably well at classifying and predicting whether a movie would win awards of some kind or not.