Mengjie Shen

Jeffrey Simonoff

Regression and Multivariate Data Analysis

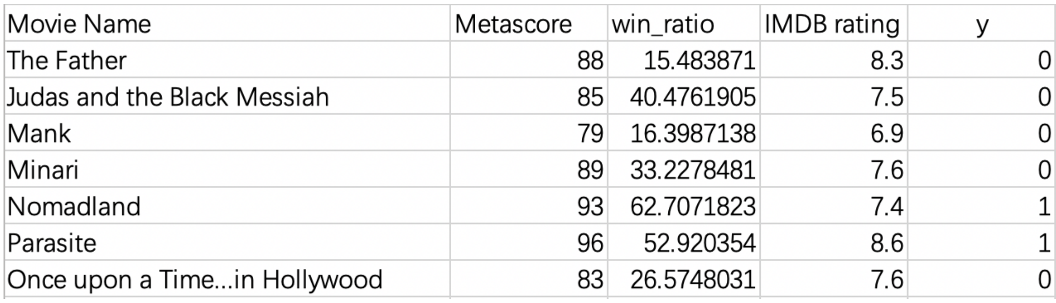
May 10, 2021

# Predicting the Oscar Award for Best Picture

Every year, the Oscar Award for Best Picture is usually regarded as the most prestigious honor of the ceremony and attracts the most attention. In my report, I would like to use several factors to distinguish the movies that won the Oscar Award and those that didn't. I chose three predicting variables which are IMDb rating, Metascore and the ratio of winning prizes. IMDb rating can represent the preference of the mass; Metascore can represent the preference of the leading critics. For the ratio of winning prizes, I use the formula of the total number of prizes that a movie won divided by the total number of prizes that a movie won or get nominated to get the ratio ($\frac{\#win}{\#win+\#nominated}$). The reason I chose this formula is that this probably can tell me the chances a movie can win the prize given that it has been nominated.

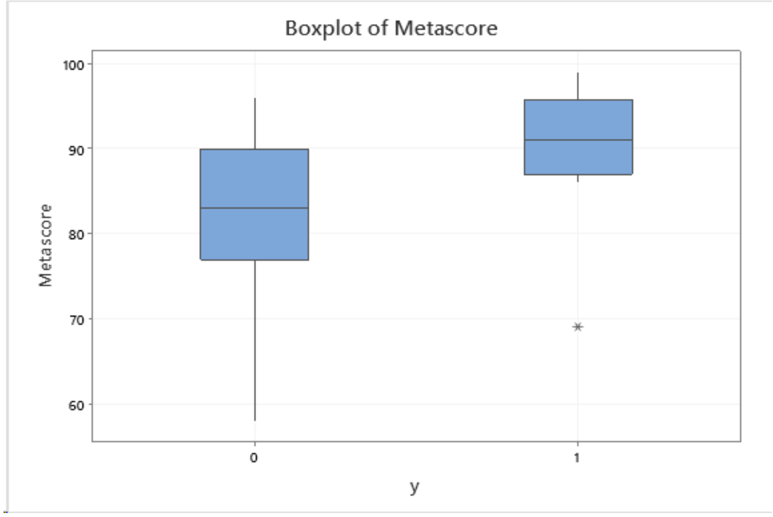
I obtained the list of nominees and winners from the database on the official website of the Academy Award (http://awardsdatabase.oscars.org/). I chose 12 movies that won the prize in the last 12 years and I chose 21 movies in the past three years that got nominated in order to make my data more balanced. Then, I obtained the corresponding IMDb Rating, Metascore and the total number of prizes it won and nominated from IMDb (https://www.imdb.com/).

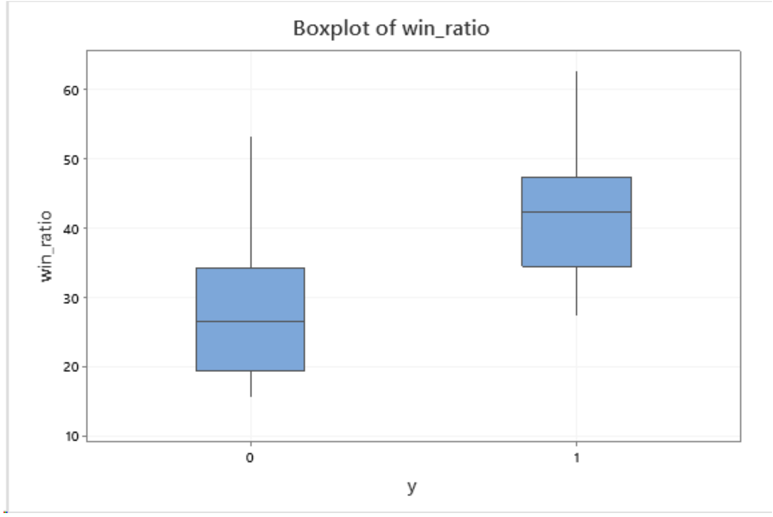
This is the head of my data:

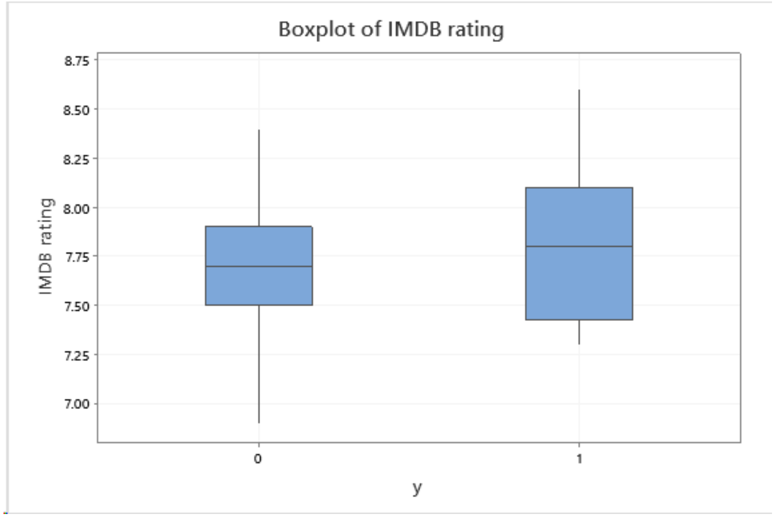


| Movie Name | Metascore | win_ratio | IMDB rating | y |
|---|---|---|---|---|
| The Father | 88 | 15.483871 | 8.3 | 0 |
| Judas and the Black Messiah | 85 | 40.4761905 | 7.5 | 0 |
| Mank | 79 | 16.3987138 | 6.9 | 0 |
| Minari | 89 | 33.2278481 | 7.6 | 0 |
| Nomadland | 93 | 62.7071823 | 7.4 | 1 |
| Parasite | 96 | 52.920354 | 8.6 | 1 |
| Once upon a Time...in Hollywood | 83 | 26.5748031 | 7.6 | 0 |

First, we construct the boxplot for each predictor:

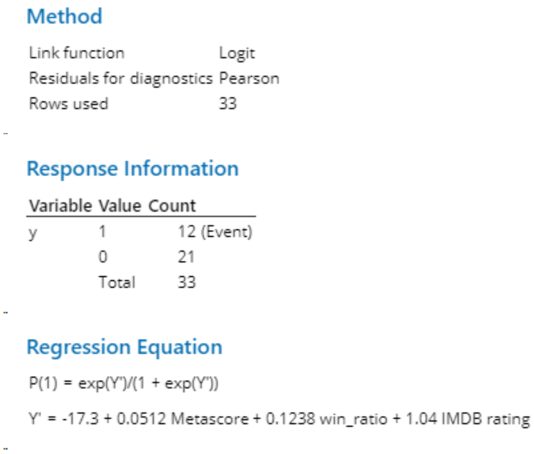




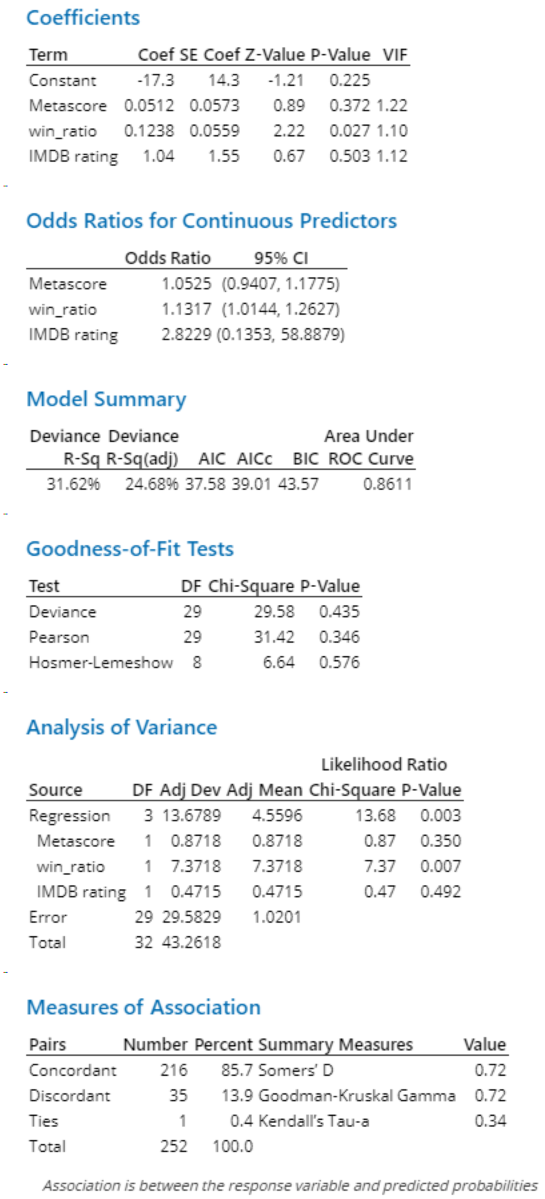


There is not a clear separation between films that won Oscar Award for the Best Picture and those that did not, especially for the Metascore and IMDB ratings. There seems to be a right-tailed pattern in the boxplot of the ratio of winning. I tried logging this variable, but the performance is quite similar, so I won't pursue it further.

Then I run the logistic regression to analyze the relationship between the predicting variables and the probability of a film that gets the Oscar Award for Best Picture more precisely. The indicator of whether the film gets the Oscar Award for Best Picture is the target variable (represents as y in Minitab). Here is the output for a logistic regression model fit to these data.



```
Method

Link function              Logit
Residuals for diagnostics  Pearson
Rows used                  33

Response Information

Variable Value Count
y        1     12 (Event)
         0     21
         Total 33

Regression Equation

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

Y' = -17.3 + 0.0512 Metascore + 0.1238 win_ratio + 1.04 IMDB rating
```



Coefficients

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | -17.3 | 14.3 | -1.21 | 0.225 | |
| Metascore | 0.0512 | 0.0573 | 0.89 | 0.372 | 1.22 |
| win_ratio | 0.1238 | 0.0559 | 2.22 | 0.027 | 1.10 |
| IMDB rating | 1.04 | 1.55 | 0.67 | 0.503 | 1.12 |

Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| Metascore | 1.0525 | (0.9407, 1.1775) |
| win_ratio | 1.1317 | (1.0144, 1.2627) |
| IMDB rating | 2.8229 | (0.1353, 58.8879) |

Model Summary

| Deviance | Deviance | | | | Area Under |
|---|---|---|---|---|---|
| R-Sq | R-Sq(adj) | AIC | AICc | BIC | ROC Curve |
| 31.62% | 24.68% | 37.58 | 39.01 | 43.57 | 0.8611 |

Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 29 | 29.58 | 0.435 |
| Pearson | 29 | 31.42 | 0.346 |
| Hosmer-Lemeshow | 8 | 6.64 | 0.576 |

Analysis of Variance

| | | | | Likelihood Ratio | |
|---|---|---|---|---|---|
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 3 | 13.6789 | 4.5596 | 13.68 | 0.003 |
| Metascore | 1 | 0.8718 | 0.8718 | 0.87 | 0.350 |
| win_ratio | 1 | 7.3718 | 7.3718 | 7.37 | 0.007 |
| IMDB rating | 1 | 0.4715 | 0.4715 | 0.47 | 0.492 |
| Error | 29 | 29.5829 | 1.0201 | | |
| Total | 32 | 43.2618 | | | |

Measures of Association

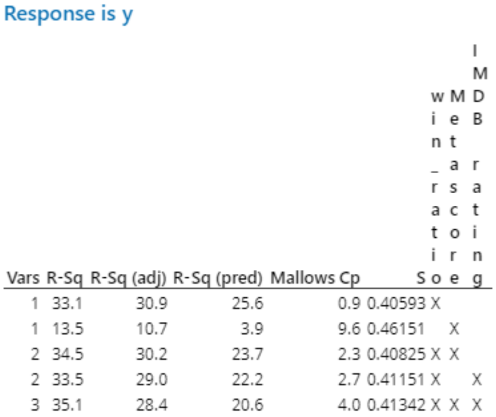
| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 216 | 85.7 | Somers' D | 0.72 |
| Discordant | 35 | 13.9 | Goodman-Kruskal Gamma | 0.72 |
| Ties | 1 | 0.4 | Kendall's Tau-a | 0.34 |
| Total | 252 | 100.0 | | |

*Association is between the response variable and predicted probabilities*

The Chi-Square of the regression is 13.68 with a p-value of 0.003. So, we strongly reject the null hypothesis of no relationship. The ratio of winning awards is highly statistically significant, while Metascore and IMDB ratings are not significant.

The coefficient for the ratio of winning awards means that holding everything else in the model fixed, an increase of one percentage point in the ratio of winning awards is associated with an increase in the odds of getting the Oscar by 13.17%. The coefficient for Metascore shows that holding everything else in the model fixed, an increase of one point in the Metascore is associated with an increase in the odds of getting the Oscar by 5.25%. The coefficient for IMDB rating shows that holding everything else in the model fixed, an increase of 0.1 in the IMDB rating is associated with multiplying the odds of getting the Oscar by 1.11 ($e^{0.1*1.04}$), which is an increase in the odds of getting the Oscar by 11%. VIF doesn't indicate any possibility of the problem of collinearity.

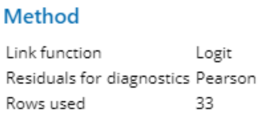
The Hosmer-Lemeshow test has a p-value equals 0.576, indicating no evidence of lack of fit. Value for Somers' D equals to 0.72 and the area under the ROC curve is 0.8611, indicating that there is good separation, with 85.7% concordant pairs and 13.9% discordant pairs. But we witness that Metascore and IMDB Ratings don't have a strong predicting power, let's perform the best subsets regression. Here is the resultant of the best subset output:



Response is y

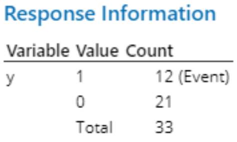
```
                                                        I
                                                        M
                                                    w M D
                                                    i e B
                                                    n t
                                                    _ a r
                                                    r s a
                                                    a c t
                                                    t o i
                                                    i r n
Vars R-Sq R-Sq (adj) R-Sq (pred) Mallows Cp       S o e g
   1 33.1       30.9        25.6        0.9 0.40593 X
   1 13.5       10.7         3.9        9.6 0.46151   X
   2 34.5       30.2        23.7        2.3 0.40825 X X
   2 33.5       29.0        22.2        2.7 0.41151 X   X
   3 35.1       28.4        20.6        4.0 0.41342 X X X
```

The best subsets point to a model with only one predictor ratio of winning awards. Let's run the logistic regression again using this variable:



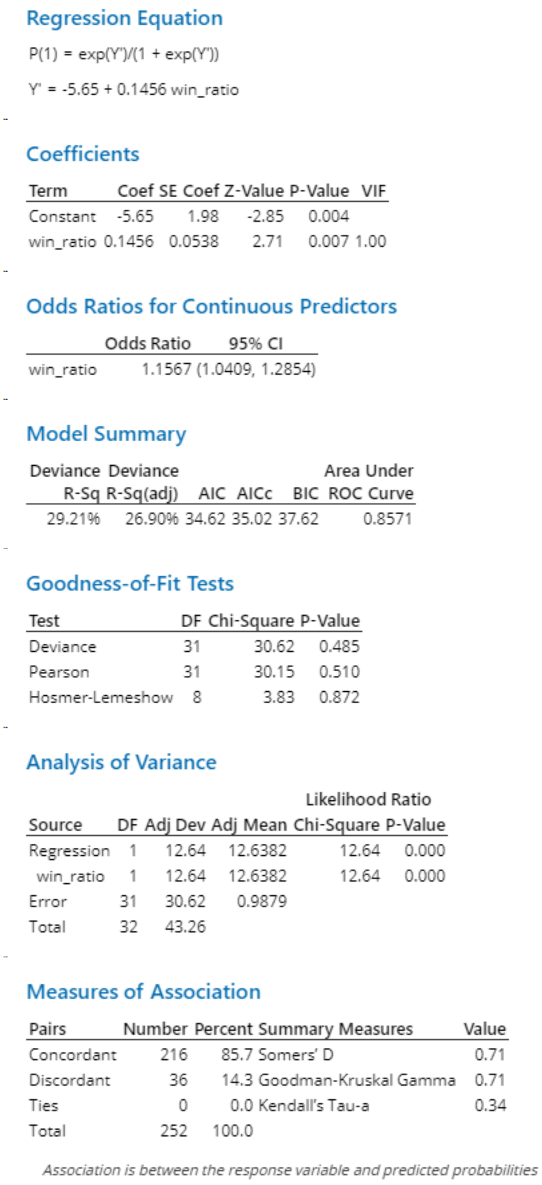
Method

```
Link function              Logit
Residuals for diagnostics  Pearson
Rows used                  33
```



Response Information

```
Variable Value Count
y        1     12 (Event)
         0     21
         Total 33
```



### Regression Equation

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

Y' = -5.65 + 0.1456 win_ratio

### Coefficients

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | -5.65 | 1.98 | -2.85 | 0.004 | |
| win_ratio | 0.1456 | 0.0538 | 2.71 | 0.007 | 1.00 |

### Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| win_ratio | 1.1567 | (1.0409, 1.2854) |

### Model Summary

| Deviance R-Sq | Deviance R-Sq(adj) | AIC | AICc | BIC | Area Under ROC Curve |
|---|---|---|---|---|---|
| 29.21% | 26.90% | 34.62 | 35.02 | 37.62 | 0.8571 |

### Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 31 | 30.62 | 0.485 |
| Pearson | 31 | 30.15 | 0.510 |
| Hosmer-Lemeshow | 8 | 3.83 | 0.872 |

### Analysis of Variance

| | | | | Likelihood Ratio | |
|---|---|---|---|---|---|
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 1 | 12.64 | 12.6382 | 12.64 | 0.000 |
| win_ratio | 1 | 12.64 | 12.6382 | 12.64 | 0.000 |
| Error | 31 | 30.62 | 0.9879 | | |
| Total | 32 | 43.26 | | | |

### Measures of Association

| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 216 | 85.7 | Somers' D | 0.71 |
| Discordant | 36 | 14.3 | Goodman-Kruskal Gamma | 0.71 |
| Ties | 0 | 0.0 | Kendall's Tau-a | 0.34 |
| Total | 252 | 100.0 | | |

*Association is between the response variable and predicted probabilities*

The $AIC_c$ value has dropped by 3.99 from the original three-predictor model, therefore, the simpler model might be a better choice. The Somers' D is 0.71, indicating that there is good separation. The Hosmer-Lemeshow test indicates no lack of fit. The predictor ratio of winning awards is statistically significant.

Next, let's identify the unusual observations which can have a strong effect on the fitted logistic regression model.

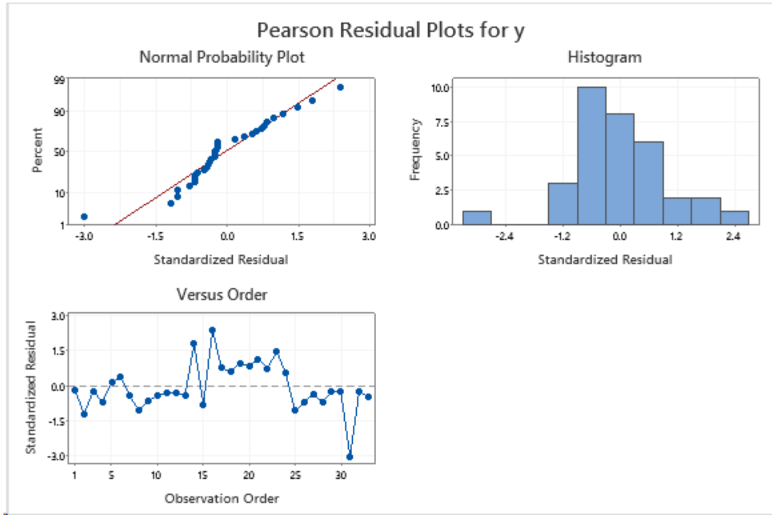
| ROWS | SPEARRES_1 | HI_1 | COOK_1 |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 1 | -0.1875 | 0.044383 | 0.000816 |
| 2 | -1.1684 | 0.063827 | 0.046537 |
| 3 | -0.20063 | 0.046468 | 0.000981 |
| 4 | -0.68326 | 0.047067 | 0.011529 |
| 5 | 0.18143 | 0.065823 | 0.00116 |
| 6 | 0.37771 | 0.103967 | 0.008277 |
| 7 | -0.42297 | 0.056059 | 0.005312 |
| 8 | -1.04728 | 0.056766 | 0.033004 |
| 9 | -0.6207 | 0.048458 | 0.00981 |
| 10 | -0.39912 | 0.05666 | 0.004784 |
| 11 | -0.26529 | 0.053768 | 0.002 |
| 12 | -0.25949 | 0.053302 | 0.001896 |
| 13 | -0.3792 | 0.056978 | 0.004344 |
| 14 | 1.79908 | 0.049647 | 0.084544 |
| 15 | -0.77781 | 0.046878 | 0.014878 |
| 16 | 2.37966 | 0.055345 | 0.165884 |
| 17 | 0.77444 | 0.077541 | 0.025207 |
| 18 | 0.6202 | 0.094796 | 0.020141 |
| 19 | 0.98763 | 0.058332 | 0.030211 |
| 20 | 0.84134 | 0.070503 | 0.026846 |
| 21 | 1.16946 | 0.049836 | 0.035866 |
| 22 | 0.74106 | 0.081279 | 0.024292 |
| 23 | 1.50028 | 0.046869 | 0.055342 |
| 24 | 0.5479 | 0.101439 | 0.016944 |
| 25 | -1.04938 | 0.056882 | 0.033208 |
| 26 | -0.6885 | 0.046995 | 0.011688 |
| 27 | -0.33841 | 0.056951 | 0.003458 |
| 28 | -0.67444 | 0.047205 | 0.011268 |
| 29 | -0.23964 | 0.051437 | 0.001557 |
| 30 | -0.22405 | 0.049665 | 0.001312 |
| 31 | -3.00875 | 0.103482 | 0.522456 |
| 32 | -0.20764 | 0.047493 | 0.001075 |
| 33 | -0.48201 | 0.053896 | 0.006618 |

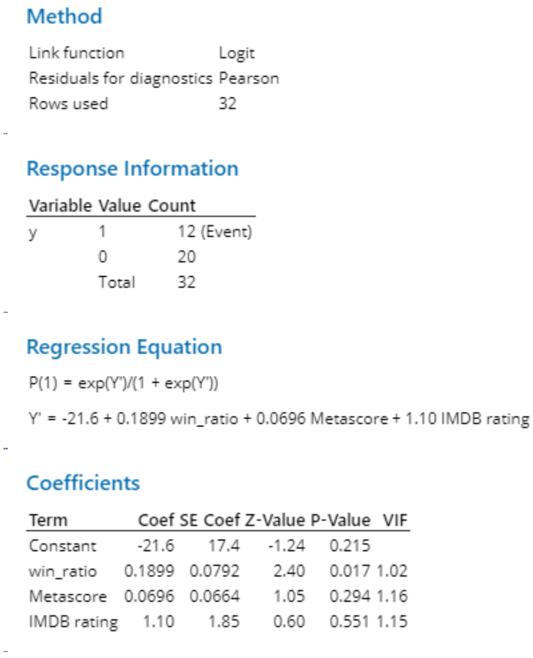
There is one outlier witnessed from row 31. It also has a relatively large Cook's Distance This corresponds to the movie Roma. It has an extremely high ratio of winning awards (53%), but it didn't get the Oscar Best Picture. However, even though it lost the biggest prize of Oscar, it actually took home three Academy Awards, including Best Cinematography, Best Foreign

Language Film and Best Director. These awards still can prove that the movie is a success. Failing to win the Oscar Best Picture can be due to some random factors.

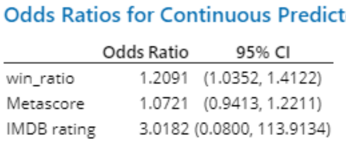
We can also witness this unusual observation in residual plots, which is at the left bottom of the normality probability plot:



Here is the fit of the three-predictor model to the dataset without the outlier:

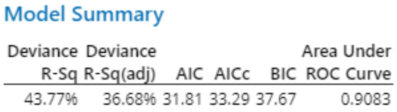


```
Method

Link function                Logit
Residuals for diagnostics    Pearson
Rows used                    32

Response Information

Variable  Value  Count
y         1      12 (Event)
          0      20
          Total  32

Regression Equation

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

Y' = -21.6 + 0.1899 win_ratio + 0.0696 Metascore + 1.10 IMDB rating

Coefficients

Term           Coef  SE Coef  Z-Value  P-Value   VIF
Constant      -21.6     17.4    -1.24    0.215
win_ratio    0.1899   0.0792     2.40    0.017  1.02
Metascore    0.0696   0.0664     1.05    0.294  1.16
IMDB rating    1.10     1.85     0.60    0.551  1.15
```



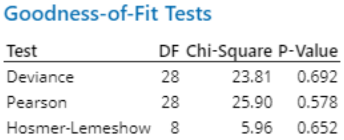
### Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| win_ratio | 1.2091 | (1.0352, 1.4122) |
| Metascore | 1.0721 | (0.9413, 1.2211) |
| IMDB rating | 3.0182 | (0.0800, 113.9134) |



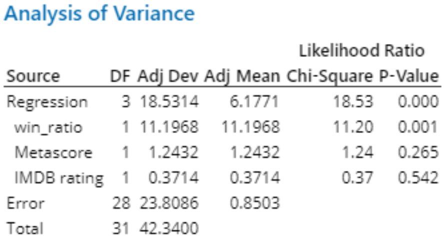
### Model Summary

| Deviance R-Sq | Deviance R-Sq(adj) | AIC | AICc | BIC | Area Under ROC Curve |
|---|---|---|---|---|---|
| 43.77% | 36.68% | 31.81 | 33.29 | 37.67 | 0.9083 |



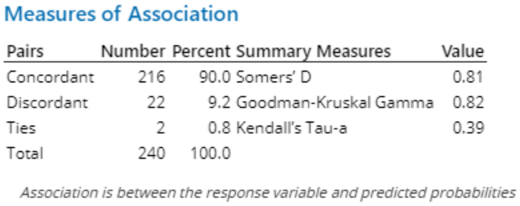
### Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 28 | 23.81 | 0.692 |
| Pearson | 28 | 25.90 | 0.578 |
| Hosmer-Lemeshow | 8 | 5.96 | 0.652 |



### Analysis of Variance

| | | | | Likelihood Ratio | |
|---|---|---|---|---|---|
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 3 | 18.5314 | 6.1771 | 18.53 | 0.000 |
| win_ratio | 1 | 11.1968 | 11.1968 | 11.20 | 0.001 |
| Metascore | 1 | 1.2432 | 1.2432 | 1.24 | 0.265 |
| IMDB rating | 1 | 0.3714 | 0.3714 | 0.37 | 0.542 |
| Error | 28 | 23.8086 | 0.8503 | | |
| Total | 31 | 42.3400 | | | |

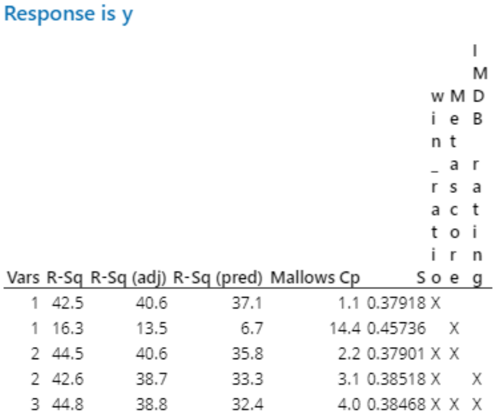


### Measures of Association

| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 216 | 90.0 | Somers' D | 0.81 |
| Discordant | 22 | 9.2 | Goodman-Kruskal Gamma | 0.82 |
| Ties | 2 | 0.8 | Kendall's Tau-a | 0.39 |
| Total | 240 | 100.0 | | |

*Association is between the response variable and predicted probabilities*

Next, we rerun the best subsets regression and here is the output:



### Response is y

| Vars | R-Sq | R-Sq (adj) | R-Sq (pred) | Mallows Cp | S | win_ratio | Metascore | IMDB rating |
|---|---|---|---|---|---|---|---|---|
| 1 | 42.5 | 40.6 | 37.1 | 1.1 | 0.37918 | X | | |
| 1 | 16.3 | 13.5 | 6.7 | 14.4 | 0.45736 | | X | |
| 2 | 44.5 | 40.6 | 35.8 | 2.2 | 0.37901 | X | X | |
| 2 | 42.6 | 38.7 | 33.3 | 3.1 | 0.38518 | X | | X |
| 3 | 44.8 | 38.8 | 32.4 | 4.0 | 0.38468 | X | X | X |

The model with only one predictor ratio of winning awards still seems to be the best model. Here is the output for the model:

## Method

| | |
|---|---|
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 32 |

## Response Information

| Variable | Value | Count |
|---|---|---|
| y | 1 | 12 (Event) |
| | 0 | 20 |
| | Total | 32 |

## Regression Equation

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

Y' = -7.83 + 0.2126 win_ratio

## Coefficients

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | -7.83 | 2.78 | -2.82 | 0.005 | |
| win_ratio | 0.2126 | 0.0769 | 2.77 | 0.006 | 1.00 |

## Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| win_ratio | 1.2369 | (1.0639, 1.4380) |

## Model Summary

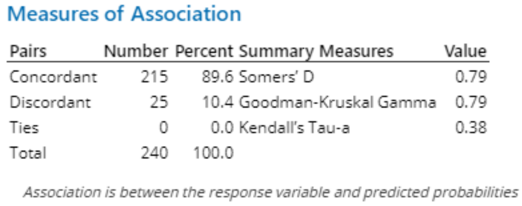
| Deviance | Deviance | | | | Area Under |
|---|---|---|---|---|---|
| R-Sq | R-Sq(adj) | AIC | AICc | BIC | ROC Curve |
| 40.74% | 38.37% | 29.09 | 29.51 | 32.02 | 0.8958 |

## Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 30 | 25.09 | 0.721 |
| Pearson | 30 | 24.20 | 0.763 |
| Hosmer-Lemeshow | 8 | 4.82 | 0.776 |

## Analysis of Variance

| | | | | Likelihood Ratio | |
|---|---|---|---|---|---|
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 1 | 17.25 | 17.2476 | 17.25 | 0.000 |
| win_ratio | 1 | 17.25 | 17.2476 | 17.25 | 0.000 |
| Error | 30 | 25.09 | 0.8364 | | |
| Total | 31 | 42.34 | | | |

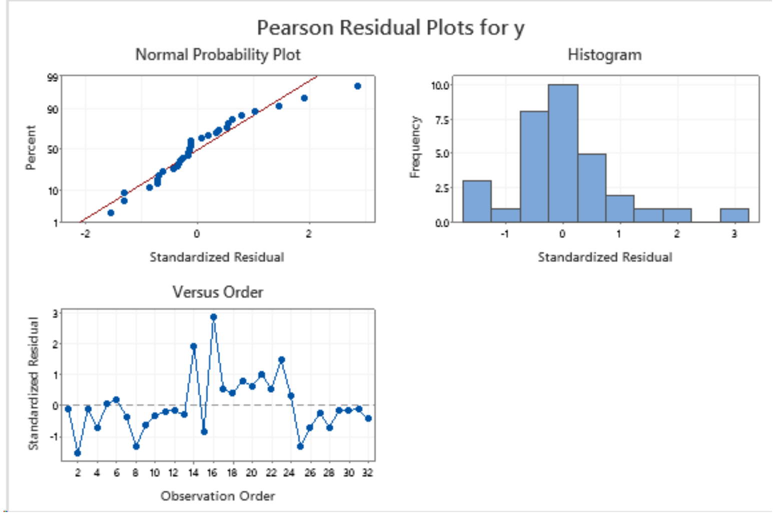


Measures of Association

| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 215 | 89.6 | Somers' D | 0.79 |
| Discordant | 25 | 10.4 | Goodman-Kruskal Gamma | 0.79 |
| Ties | 0 | 0.0 | Kendall's Tau-a | 0.38 |
| Total | 240 | 100.0 | | |

*Association is between the response variable and predicted probabilities*

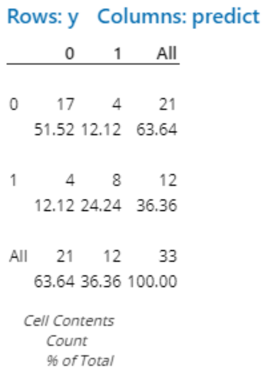
The coefficient for the ratio of winning awards means that holding everything else in the model fixed, an increase of one percentage point in the ratio of winning awards is associated with an increase in the odds of getting the Oscar by 23.69%. Here are the three-in-one plot for the model. There are still some indications of unusual observations in the model, however, since a new outlier will keep popping up if we subset the data, we will stop here.



| Movie Name | FITS | SPEARRES | HI | COOK |
|---|---|---|---|---|
| The Father | 0.032505 | -0.1875 | 0.044383 | 0.000816 |
| Judas and the Black Messiah | 0.561022 | -1.1684 | 0.063827 | 0.046537 |
| Mank | 0.036965 | -0.20063 | 0.046468 | 0.000981 |
| Minari | 0.307896 | -0.68326 | 0.047067 | 0.011529 |
| Nomadland | 0.970169 | 0.18143 | 0.065823 | 0.00116 |
| Parasite | 0.886658 | 0.37771 | 0.103967 | 0.008277 |
| Once upon a Time...in Hollywood | 0.144478 | -0.42297 | 0.056059 | 0.005312 |
| 1917 | 0.508485 | -1.04728 | 0.056766 | 0.033004 |
| Marriage Story | 0.268258 | -0.6207 | 0.048458 | 0.00981 |
| Little Women | 0.130642 | -0.39912 | 0.05666 | 0.004784 |

| | | | | |
|---|---|---|---|---|
| Ford v Ferrari | 0.062439 | -0.26529 | 0.053768 | 0.002 |
| Vice | 0.059927 | -0.25949 | 0.053302 | 0.001896 |
| A Star Is Born | 0.119408 | -0.3792 | 0.056978 | 0.004344 |
| Green Book | 0.245338 | 1.79908 | 0.049647 | 0.084544 |
| The Favorite | 0.365735 | -0.77781 | 0.046878 | 0.014878 |
| The Shape of Water | 0.157496 | 2.37966 | 0.055345 | 0.165884 |
| Moonlight | 0.643813 | 0.77444 | 0.077541 | 0.025207 |
| Spotlight | 0.741736 | 0.6202 | 0.094796 | 0.020141 |
| Birdman | 0.521234 | 0.98763 | 0.058332 | 0.030211 |
| 12 Years a Slave | 0.603156 | 0.84134 | 0.070503 | 0.026846 |
| Argo | 0.434879 | 1.16946 | 0.049836 | 0.035866 |
| The Artist | 0.664661 | 0.74106 | 0.081279 | 0.024292 |
| The King’ s Speech | 0.317929 | 1.50028 | 0.046869 | 0.055342 |
| The Hurt Locker | 0.787562 | 0.5479 | 0.101439 | 0.016944 |
| Promising Young Woman | 0.509457 | -1.04938 | 0.056882 | 0.033208 |
| Sound of Metal | 0.311178 | -0.6885 | 0.046995 | 0.011688 |
| The Trial of the Chicago 7 | 0.097473 | -0.33841 | 0.056951 | 0.003458 |
| Joker | 0.302358 | -0.67444 | 0.047205 | 0.011268 |
| Jojo Rabbit | 0.05166 | -0.23964 | 0.051437 | 0.001557 |
| The Irishman | 0.045534 | -0.22405 | 0.049665 | 0.001312 |
| Roma | 0.8903 | -3.00875 | 0.103482 | 0.522456 |
| BlacKkKlansman | 0.039447 | -0.20764 | 0.047493 | 0.001075 |
| Black Panther | 0.180202 | -0.48201 | 0.053896 | 0.006618 |

Next, we put the unusual observation back to our dataset and plot the classification matrix.



```
Rows: y   Columns: predict

          0      1    All

0        17      4     21
      51.52  12.12  63.64

1         4      8     12
      12.12  24.24  36.36

All      21     12     33
      63.64  36.36 100.00

  Cell Contents
     Count
     % of Total
```

75.76% of the films were correctly classified, much higher than

$$C_{pro} = 1.25 \times (0.6364 \times 0.6364 + 0.3636 \times 0.3636) = 67.15\%$$

And $C_{max} = 63.64\%$. Therefore, the model performs well in predicting which movie can win Oscar Award for Best Picture.

Next, we would like to adjust the intercept term in order to estimate the prospective probabilities of getting an Oscar Award for Best Picture. Since the winner is usually generated from 8 to 9 nominees, I will use a 12.5% prior probability of winning the award. Then, the adjusted intercept is:

$$\widetilde{\beta_0} = \widehat{\beta_0} + ln\left[\frac{0.125 \times 20}{0.875 \times 12}\right] = -9.27$$

I can then convert the original probability estimates to adjusted ones, and here is the estimated probability:

| row | Movie Name | newprob |
|---|---|---|
| 1 | The Father | 0.008217 |
| 2 | Judas and the Black Messiah | 0.239633 |
| 3 | Mank | 0.009377 |
| 4 | Minari | 0.098858 |
| 5 | Nomadland | 0.889133 |
| 6 | Parasite | 0.658598 |
| 7 | Once upon a Time...in Hollywood | 0.03998 |
| 8 | 1917 | 0.203258 |
| 9 | Marriage Story | 0.082908 |
| 10 | Little Women | 0.035733 |
| 11 | Ford v Ferrari | 0.016157 |
| 12 | Vice | 0.015476 |
| 13 | A Star Is Born | 0.032356 |
| 14 | Green Book | 0.074218 |
| 15 | The Favorite | 0.124492 |
| 16 | The Shape of Water | 0.044067 |
| 17 | Moonlight | 0.308307 |
| 18 | Spotlight | 0.414599 |
| 19 | Birdman | 0.211649 |
| 20 | 12 Years a Slave | 0.27262 |
| 21 | Argo | 0.159498 |
| 22 | The Artist | 0.328304 |

| | | |
|---|---|---|
| 23 | The King’s Speech | 0.103094 |
| 24 | The Hurt Locker | 0.477589 |
| 25 | Promising Young Woman | 0.203888 |
| 26 | Sound of Metal | 0.100235 |
| 27 | The Trial of the Chicago 7 | 0.025942 |
| 28 | Joker | 0.096556 |
| 29 | Jojo Rabbit | 0.013255 |
| 30 | The Irishman | 0.011627 |
| 31 | Roma | 0.666815 |
| 32 | BlacKkKlansman | 0.010025 |
| 33 | Black Panther | 0.051418 |

There are 7 movies that are classified incorrectly (Judas and the Black Messiah, 1917, Green Book, The Shape of Water, The King’s Speech, Promising Young Woman, Roma). Notably, Roma has an estimated probability of 66.68% of getting the Oscar Awards, lower than the previous one 89.03%, but it is still misclassified.

To conclude, the model with only one predictor of the ratio of winning awards is the best model. I was hoping for a stronger relationship between either Metascore or IMDb rating because this represents the taste of professional critics and the audience, which should be consistent with the result of the Oscar Award for the Best Picture. The ratio of winning the awards is quite consistent with the result of the Oscar Award. I think this probably because the result of other awards can reflect the view of the critics in the industry to some extent. In the future, more variables such as the box office or movie genre can also be considered into the model to see whether there can be a more accurate prediction.