Academy Awards: Modelling and Prediction

MATH 396 Midterm Report

Christopher Lee

christopher.lee2@mail.mcgill.ca

March 1, 2014

Table of Contents

Introduction I

The Academy Awards represent the ultimate culmination of a film's critical success. It is the final and most important film award in the award season for the industry of motion picture. Studies have even suggested (contentiously) that Oscar winners experience increased life expectancy. The Oscars represent a huge financial undertaking by film studios and producers for big-budget awards-campaigning. Also, prediction markets are trading millions of dollars in Oscar betting.

Introduction II

It is my intention to hollistically gather data on critically acclaimed and Oscar nominated films in order to model and predict the outcome of the annual Academy Awards in six categories.

- Best Actor in a Leading Role
- Best Actress in a Leading Role
- Best Actor in a Supporting Role
- Best Actress in a Supporting Role
- Best Directing
- Best Picture

Introduction III

The first goal of this project is predictive modelling. I will endeavor to find models that best estimate the odds of Ocsar nominees winning. I will check the models for fit, and accuracy of prediction.

The second goal is descriptive modelling. Here we will focus more on the relationships between the variables correlated with the odds of winning an oscar, how they change across category and how they interact with other variables. I will scrutinize for spurious relationships and confounder variables

Section 1

Data Collection and Webscrapping

The Data

There is a stark lack of clean datasets, or data-friendly spreadsheets available for film. Therefore a large aspect of this research has committed to creating code to scrape and create the first holistic dataset on Academy Awards. The dataset will later be released onto github.com and other data-propogating sources for further analysis by others.

Data Sources I

To begin, I employ a web-scraper written exclusively in R's RcurI CITE and XML CITE packages. The web-scraper will sift htmlTable environments and individual XM elements from the following websites

1. `imdb.com`
   The main source of data with data on film awards and major film characteristics
2. `boxofficemojo.com`
   The secondary source with reliable data on the finances of film

Data Sources II

3. `www.the-numbers.com/movie/budgets/all`
   A supplementary financial data source

4. `nndb.com`
   The bibliographical data source for actors/actresses/directors

5. `metacritic.com`
   An aggregate website which quantifies film quality on weighted average of aggregate reviews. Metacritic score will be used as a proxy for the critical reception of films.

Web-scrapper

The Web-scrapper scrapes data from a total of 4826 webpages, returning 1343 observations across 44 years (1970-2013) and 5 competitive Oscar categories. We have 37 attributes for every row.

The code for the web-scrapper itself will be made available in a separate .R file.

Covariates I

Name	Description: (C)count (B)binary (c)continuous
past.win	(C) Previous Oscars won
past.nom	(C) Previous Oscar nominations
other.wins	(C) Other awards by film
other.noms	(C) Other nominations by film
domestic.gross	(c) US Gross Earnings per million
metacritic	(c) Metacritic score

Covariates II

Name	Description: (C)count (B)binary (c)continuous
globes .	(B) Won 2014 Golden Globes award in same category
bafta	(B) Won 2014 BAFTA in same category
dga	(B) Won 2014 Directors Guild Award
sag	(B) Won 2014 Screen Actors Guild Award
adapted	(B) Film adapted from another medium
date	(c) Month of film's wide release

Covariates III

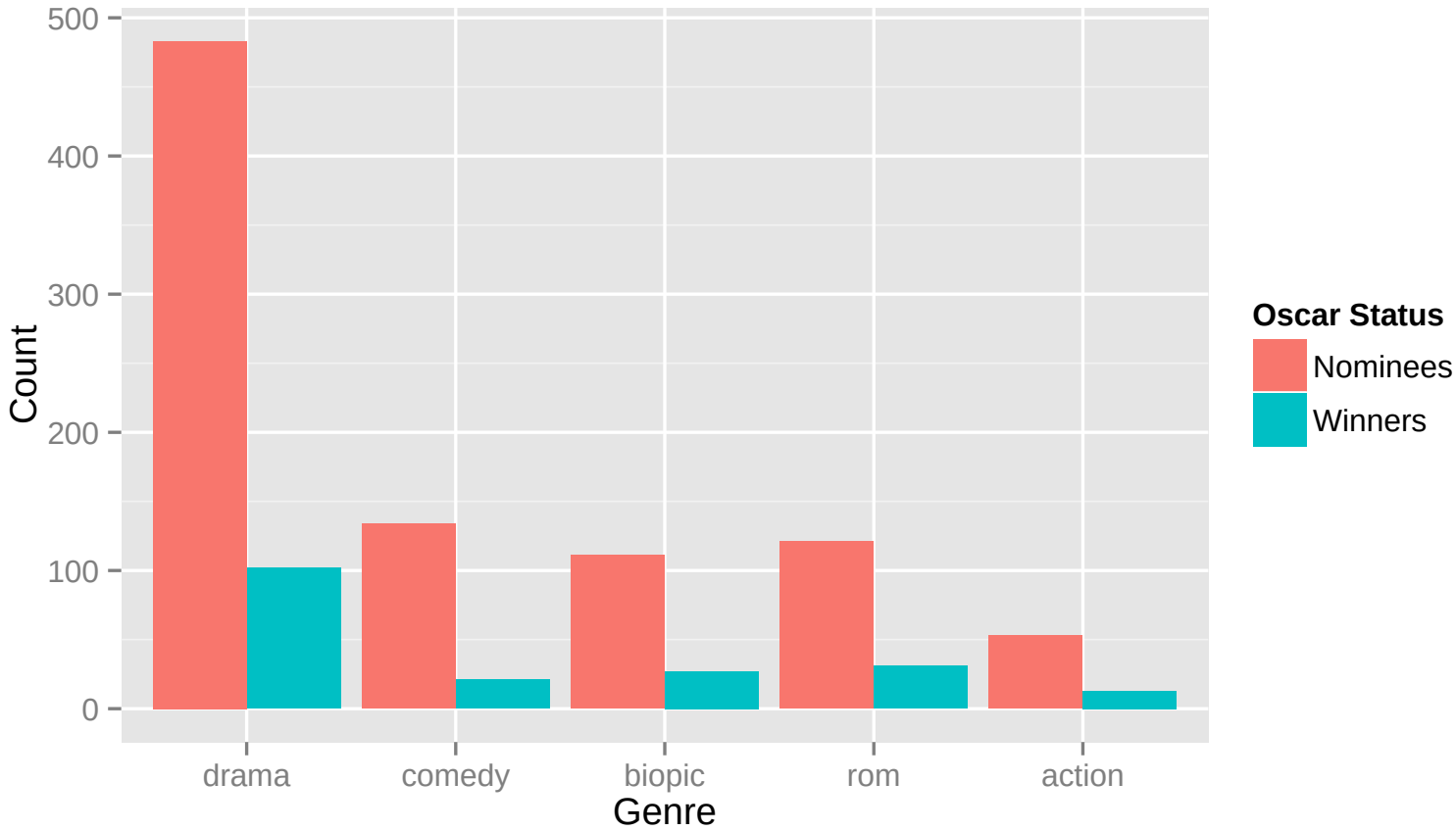
Name	Description: (C)count (B)binary (c)continuous
picture.nom	(B) Oscar Nomination for Best Picture
direct.nom	(B) Oscar Nomination for Best Director
edit.nom	(B) Oscar Nomination for Best Editing
script.nom	(B) Oscar Nomination for Best Screenplay (adapted or original)
.	
tiff.premiere	(B) Film Premiere at Toronto International Film Festival

Section 3

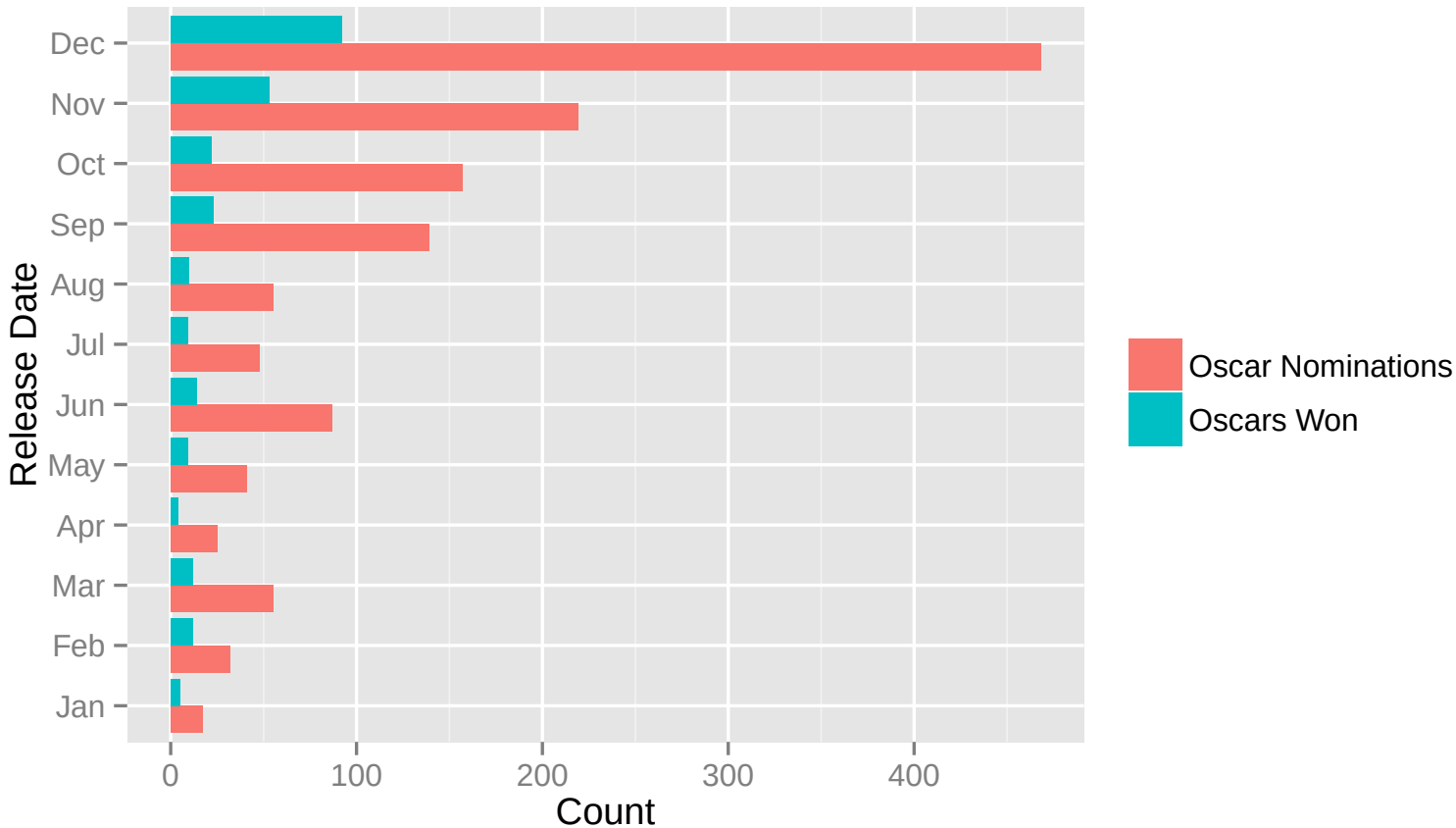
Exploratory Data Analysis

Convention wisdom suggests some characteristics about the Oscar ceremony. We will quantitatively verify the claims of these expert pundits.

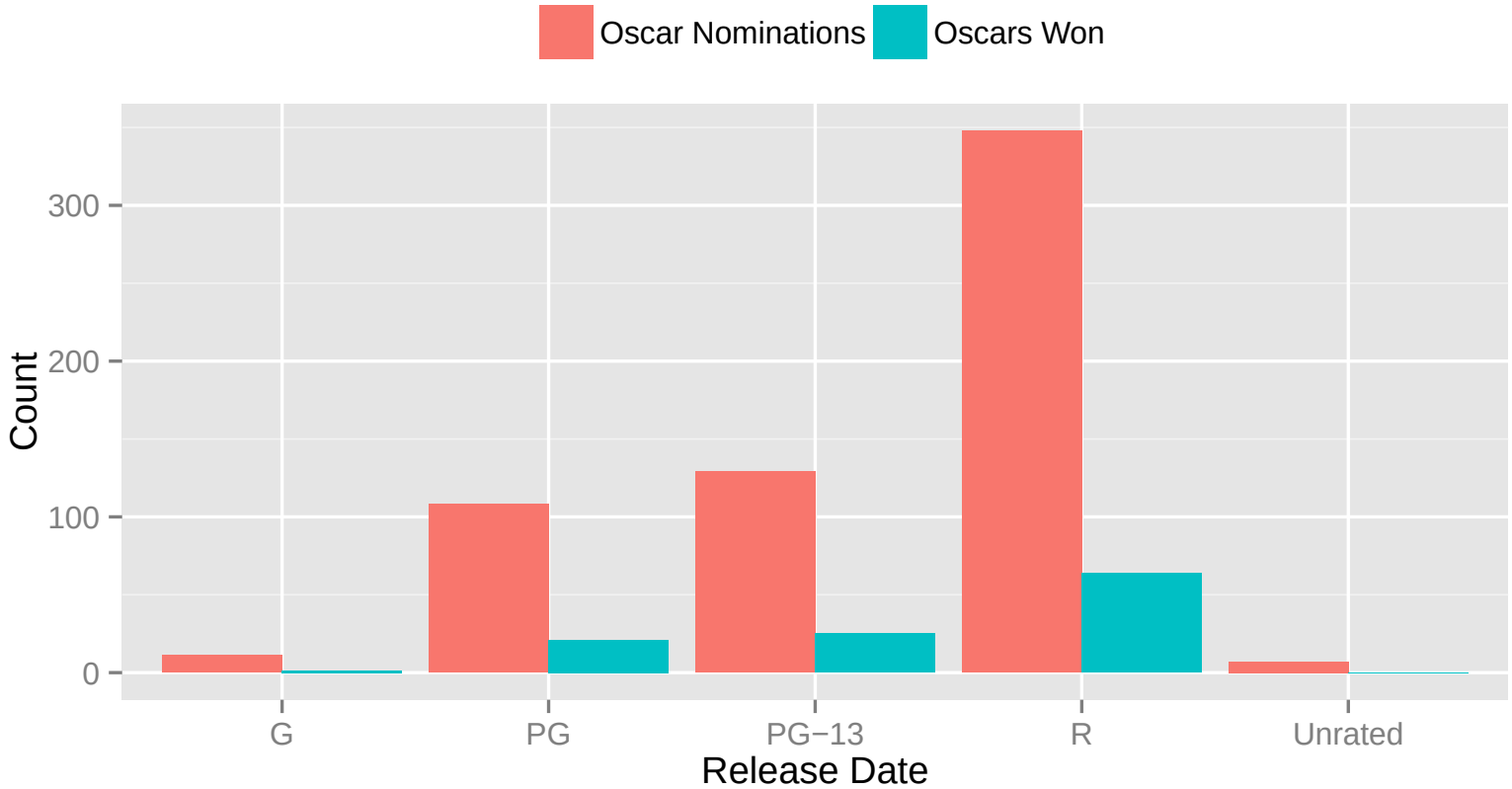
Genre Discrimination



Release Date trends



The R-rated Academy?



An interesting avenue to approach is the idea that the Academy endeavors to reward so-called 'high-art' or cinematic projects that are mature and uncomfortable/inappropriate for younger audiences.

Genres and Categories

Category	drama	comedy	biopic	rom	action	adapted	age
Best Actor	0.94	0.20	0.28	0.18	0.07	0.60	46.50
Best Actress	0.93	0.21	0.19	0.34	0.03	0.61	39.54
Best Director	0.91	0.19	0.24	0.20	0.14	0.56	49.65
Best Picture	0.90	0.16	0.26	0.23	0.15	0.57	
Best Supporting Actor	0.87	0.21	0.22	0.16	0.13	0.57	49.08
Best Supporting Actress	0.89	0.31	0.15	0.30	0.04	0.63	40.51

It is not surprising that to see that Best Picture holds the comedy genre in the lowest regard as seen by the meek representation, and favors biographical feature films. Actress nominees have the lowest mean age while directors have the highest. We also see that over half of all nominated films already exist in some other medium as 59% of all nominees are adapted from other sources.

Most decorated Winners and Nominees

The top 5 most nominated individuals

Actors		**Actresses**		**Directors**	
Name	Nominations	Name	Nominations	Name	Nominations
Jack Nicholson	12	Meryl Streep	17	Martin Scorsese	7
Al Pacino	8	Jane Fonda	7	Steven Spielberg	7
Robert De Niro	7	Sissy Spacek	7	Woody Allen	7
Denzel Washington	6	Ellen Burstyn	6	Robert Altman	5
Dustin Hoffman	6	Glenn Close	6	Clint Eastwood	4

Now the top 5 winners

Actors		**Actresses**		**Directors**	
Name	Won	Name	Won	Name	Won
Daniel Day-Lewis	3	Meryl Streep	3	Ang Lee	2
Jack Nicholson	3	Dianne Wiest	2	Clint Eastwood	2
Christoph Waltz	2	Glenda Jackson	2	Milos Forman	2
Denzel Washington	2	Hilary Swank	2	Oliver Stone	2
Dustin Hoffman	2	Jane Fonda	2	Steven Spielberg	2

Section 3

Methodology: Logistic Regression

The dataset is composed of all Oscar nominees in the past 44 years. I intend to model the outcome of six award categories. The regressand, titled 'Won' is a categorical 0/1 variable. We will employ the logistic regression classification method to model the outcome. We have a modest sample size (n=220) for each category, which we will model seperately. We will apply the same model for all four acting categories, and seperate models for Best Director and Best Picture, respectively.

Probabilities and Odds

In logistic regression we are regressing covariates on a categorical variable. Our regressand Y takes values of 0 and 1.

- p denotes the *probability* of an event occuring.
- $\frac{p}{1-p}$ is the *odds* of that event occuring.
- $\ln(\frac{p}{1-p})$ is the natural logarithm of the odds, or the *logit*

$$y_i = \begin{cases} 1 & \text{if nominee has won} \\ 0 & \text{otherwise} \end{cases}$$

$$\Pr(Y_i = 1) = p_i$$

$$y_i \sim \text{Bernouilli}(p_i)$$

$$\text{odds}(Y_i = 1) = \frac{p_i}{1-p_i}$$

Logistic Regression: Linear vs. Logistic

In linear regression, our covariates have a direct linearly relationship with the regressand, but for logistic regression, the covariates have a linear relationship with the logit of the regressand.

$$y = \alpha + \beta X + \epsilon$$

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X + \epsilon$$

Assumptions

1. Observations are independent
2. Covariates are linearly related to the logit of the dependent
3. Absence of multicollinearity

Unlike OLS, logistic regression does not require a linearly relationship between the dependent and the covariates. There is no distribution assumption over variables and there is no homoskedastic assumptions being made.

Interpretation

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

We interpret the β coefficients in two equivalent ways

1. A 1-unit change in x_1 will lead to a β_1 increase in the log odds of y
2. A 1-unit change in x_1 will change the odds of y by a factor of e^{β_1}

Interpretation

Interpretation 1 follows strictly from the formula.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Interpretation 2 comes from exponentiating the formula.

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon)$$

Interpretation 2 is easier to communicate so I will predominantly report results in the exponentiated form. e^{β}'s are called an *Odds ratios*

Odds Ratios

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

$$= e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2}$$

$$= (\mathrm{OR}_0)(\mathrm{OR}_1{}^{x_1})(\mathrm{OR}_2{}^{x_2})$$

where $\mathrm{OR}_i = e^{\beta_i}$

$\frac{p}{1-p}$ and OR_i have a *multiplicative* relationship instead of the *additive* relationship between y and β_i in the OLS case.

So a 1-unit increase in x_1 changes the odds of y by a factor of OR_1 but a 2-unit increase in x_1 changes the odds by a factor of $\mathrm{OR}_1{}^2$, not $2 \times \mathrm{OR}_1$

Predicted Values

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2$$

$$\frac{\hat{p}}{1-\hat{p}} = \exp(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2)$$

$$\hat{p} = \frac{\exp(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2)}{1 + \exp(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2)}$$

This returns predicted probabilities for our regressand Y

Awards Races I

Because this analysis is not motivated or backed by any formal theory, as would be the case with an epidemiology study, we will rely on facts and trends that are widely agreed upon in the film and critic community.

1. The Screen Actors Guild Awards predict the Oscar Acting awards with great success
2. The Directors Guild awards predict the Oscar Directing award with great success
3. The British Academy of Film and Arts (BAFTA), Golden Globes and Toronto International Film Festival (TIFF) are indicative of Oscar chances

Awards Races II

4. The Academy Awards are attracted to commercially succesful projects

5. It is nearly impossible to win Best Picture without nominations for Best Editing and Best Direction.

I will first look at the Acting race.

Acting Models I with raw coefficients

	Dependent variable:			
	Won Oscar			
	Best Actor	Best Actress	Best Supporting Actor	Best Supporting Actress
globes	2.334***	2.138***	2.631***	1.871***
	(0.456)	(0.478)	(0.439)	(0.439)
bafta	1.883***	1.341***	0.420	1.644***
	(0.540)	(0.511)	(0.582)	(0.513)
picture.nom	0.979*	1.071**	0.591	1.139**
	(0.514)	(0.495)	(0.449)	(0.484)
domestic.gross	0.008*	0.003	0.001	-0.006
	(0.004)	(0.003)	(0.003)	(0.005)
past.win	-1.023**	-1.164**	-0.452	-0.390
	(0.517)	(0.561)	(0.725)	(0.719)
past.nom	0.235*	0.263**	0.149	-0.092
	(0.123)	(0.120)	(0.157)	(0.235)
Constant	-3.834***	-3.019***	-2.775***	-2.388***
	(0.581)	(0.442)	(0.436)	(0.388)
Observations	197	187	205	197
Log Likelihood	-66.390	-67.310	-77.740	-80.110
Akaike Inf. Crit.	146.800	148.600	169.500	174.200
Note:				*p<0.1; **p<0.05; ***p<0.01

β
(se)

Acting Models I

The Golden Globes and BAFTAs show the co-movement between film awards, while the picture.nom suggests that voters may be less likely to value a performance, if the film itself is poor. Also, interestingly, the log odds increase with the more previous nominations of a nominee, but decreases for every additional previous win (and we know no Actor or Director has ever won more than 3)
There is also some unanticipated features. I find the sample populations of these four categories to be disimilar in many more ways than anticipated. Most covariates are not stable across all 4 models. The logit of all four models are positively correlated with a win at the BAFTA Awards or the Golden Globes but beyond that, we can little generalize across all 4 categories.

I will examine the odds ratios of these models for further interpretation

Acting Models expressed in Odds Ratios

	Dependent variable:			
	Won Oscar			
	Best Actor	Best Actress	Best Supporting Actor	Best Supporting Actress
globes	10.310***	8.487***	13.880***	6.495***
	(4.314,26.090)	(3.386,22.380)	(6.022,34.020)	(2.791,15.770)
bafta	6.574***	3.824***	1.522	5.176***
	(2.327,19.610)	(1.400,10.550)	(0.471,4.676)	(1.900,14.410)
picture.nom	2.662*	2.917**	1.807	3.125**
	(1.000,7.662)	(1.106,7.807)	(0.753,4.463)	(1.236,8.331)
domestic.gross	1.008*	1.003	1.001	0.994
	(1.000,1.016)	(0.996,1.009)	(0.995,1.007)	(0.984,1.002)
past.win	0.359**	0.312**	0.636	0.677
	(0.122,0.950)	(0.097,0.891)	(0.139,2.459)	(0.152,2.630)
past.nom	1.265*	1.301**	1.160	0.912
	(0.982,1.605)	(1.031,1.652)	(0.840,1.565)	(0.535,1.365)
Constant	0.022***	0.049***	0.062***	0.092***
	(0.006,0.061)	(0.019,0.109)	(0.025,0.138)	(0.041,0.188)
Observations	197	187	205	197
Log Likelihood	-66.390	-67.310	-77.740	-80.110
Akaike Inf. Crit.	146.800	148.600	169.500	174.200
Note:				*p<0.1; **p<0.05; ***p<0.01

e^{β}
(C.I.)

The Acting Races

These models are only preliminary but they do show the sheer predictive power of the Golden Globes and BAFTA award shows that precede the Oscar ceremony. With 95% confidence, a Leading Actor win at the Golden Globes could raise an Oscar nominees odds, anywhere from 4 to 26 times its previous value! (holding other variables constant, of course). The story is even more drastic for Supporting Actors who have an odds ratio of 13.88 for the Globes with a 95% confidence interval of [6, 34]

Acting Models II

Notably absent from our first Acting Models are the the results of the Screen Actors Guild Awards. Experts view the SAG as the single most critical moment in determining an Academy Award winner. For our purposes, the SAG awards did not begin until 1995, thus including it in the model imposes a very large penalty to our sample size, which is not enormous to begin with.

Acting Models II expressed in Odds Ratios

	Dependent variable:			
	Won Oscar			
	Best Actor	Best Actress	Best Supporting Actor	Best Supporting Actress
sag	87.200***	35.290***	10.700***	2.059
	(12.210,1,835.000)	(8.146,215.400)	(2.670,49.750)	(0.440,9.753)
bafta	0.656	4.787*	3.086	33.340***
	(0.029,5.730)	(0.928,28.560)	(0.551,16.770)	(6.556,267.000)
globes	2.922	7.706**	12.840***	14.300***
	(0.455,16.810)	(1.618,44.510)	(3.234,59.120)	(2.663,113.800)
Constant	0.046***	0.027***	0.044***	0.026***
	(0.013,0.121)	(0.005,0.087)	(0.012,0.114)	(0.004,0.085)
Observations	95	95	96	95
Log Likelihood	-23.130	-25.800	-29.190	-25.290
Akaike Inf. Crit.	54.250	59.610	66.380	58.570
Note:				*p<0.1; **p<0.05; ***p<0.01

Acting Models II

The SAG's effect on the Oscar odds are interesting. Many of our strongest covariates in our previous models can no longer change the mean log odds for any reasonable confidence level. This is a puzzling feature we will challenge later. These results are supsect to inadequate sample size and multicollinearity, even though my VIF tests did not show it.

The results of these alternative models are confusing. An Oscar contender has next to no chance of winning the acting awards without the Screen Acting Guild nod, but *only* if he/she is in the running for the *Lead* award. But it is not signifcant for Supporting Actress! You can see that 1 falls in the bounds of the 95% confidence interval (β=0). For Supporting Actor, it is a significant predictor, but does not affect the odds as much as the Golden Globes.

The race for Director and Picture

Best Director and Best Picture are the two most closely tied categories at the Oscars. Only 4 films in history have won Best Picture without a Best Director nomination. At the other end, exactly 0 films have won Best Director without a Best Picture nomination.

At the same time, there are differences. Direction, like acting is a honed and specific craft while Best Picture is a general claim on the 'best' film. We will try to model these races.

Director and Picture Models in Odds Ratios

	Dependent variable:	
	Won Oscar	
	Best Director	Best Picture
globes	5.038***	3.936***
	(2.136,12.170)	(1.632,9.645)
bafta	2.906**	4.591***
	(1.074,7.889)	(1.808,12.130)
domestic.gross	1.006**	1.004
	(1.001,1.012)	(0.999,1.010)
edit.nom	4.958***	8.176***
	(1.711,18.130)	(2.546,37.280)
script.nom	50.130**	15.760*
	(2.913,4,116.000)	(1.318,900.100)
direct.nom		4.461*
		(1.077,31.750)
Constant	0.001***	0.0003***
	(0.00000,0.014)	(0.00000,0.007)
Observations	208	227
Log Likelihood	-75.480	-75.240
Akaike Inf. Crit.	163.000	164.500
Note:	*p<0.1; **p<0.05; ***p<0.01	

More surprises occur here. Despite close ties with Best Director, both the Director and Picture races' odds are heavily influenced by the Oscar Editing nominations by an incredible factor of OR=50 and OR=10 respectively. It also seems that the ties between Picture and Director are not as strong as first thought.

Director and Picture races

This may not provide us with a useful prediction. The editing nomination has an extreme and likely mispecified effect on the log odds of winning Best Picture and Director, but most contenders in these categories will have editing nominations. Meaning, we may be predicting several nominees with probabilities of winning in the upper 90-ith percentile.

We also know, like with the SAG awards, the DGAs may provide us a very strong predictor.

Director Model II

	Dependent variable:		
	Won Oscar		
globes	5.038***	0.864	
	(2.136,12.170)	(0.112,4.349)	
bafta	2.906**	2.354	
	(1.074,7.889)	(0.422,11.710)	
domestic.gross	1.006**	1.001	
	(1.001,1.012)	(0.993,1.009)	
edit.nom	4.958***	5.371*	5.542**
	(1.711,18.130)	(0.960,42.560)	(1.133,36.380)
script.nom	50.130**	5.948	
	(2.913,4,116.000)	(0.369,947.600)	
dga		240.400***	288.100***
		(52.540,1,925.000)	(76.850,1,548.000)
Constant	0.001***	0.002***	0.011***
	(0.00000,0.014)	(0.00001,0.041)	(0.002,0.043)
Observations	208	208	221
Log Likelihood	-75.480	-33.340	-34.790
Akaike Inf. Crit.	163.000	80.690	75.580
Note:			*p<0.1; **p<0.05; ***p<0.01

This is a result similar to including SAG into our nested Acting models. The effects of globes and bafta and others are much reduced when DGA is introduced.

The DGA has the benefit of retaining our sample size (unlike the SAG). Before more rigorous model selection, we already see a reduction in AIC and a higher log likelihood than in the previous model.

Recall, we believed that a Director's Nomination to be a significant predictor for the Best Picture race. And we also know the DGA is highly correlated with the Director Nomination. Now we have reason to entertain the possibility that the DGA win is the real significant predictor, and Direction nomination has only a spurious correlation with winning the Best Picture race.

Picture Model II

	Dependent variable:		
	Won Oscar		
globes	3.936***	1.397	
	(1.632,9.645)	(0.337,5.289)	
bafta	4.591***	4.343**	5.995***
	(1.808,12.130)	(1.153,17.080)	(1.559,23.640)
domestic.gross	1.004	1.001	
	(0.999,1.010)	(0.994,1.009)	
edit.nom	8.176***	8.893**	12.100***
	(2.546,37.280)	(1.755,65.380)	(2.598,82.750)
script.nom	15.760*	4.560	
	(1.318,900.100)	(0.283,486.900)	
direct.nom	4.461*	4.061	
	(1.077,31.750)	(0.531,51.770)	
dga		70.430***	101.600***
		(21.640,289.300)	(33.290,382.100)
Constant	0.0003***	0.0004***	0.004***
	(0.00000,0.007)	(0.00000,0.015)	(0.001,0.022)
Observations	227	227	239
Log Likelihood	-75.240	-42.200	-44.090
Akaike Inf. Crit.	164.500	100.400	96.190
Note:			*p<0.1; **p<0.05; ***p<0.01

Now, a spurious relationship between direct.nom and our dependent seems more likely. Direct.nom seems to have been masking the confounding variable: DGA

The Guild factor

Interestingly, we expected all 5 categories to follow their respective guild awards. For acting, it is the Screen Actors Guild Awards and for directors, it is the Directors Guild Awards. Modelling the Actors and Directors without the SAG and DGA results respectively, we find several covariates significantly different from 0, most notably, the respective Golden Globes and BAFTA awards contribute positively and strongly to the logit of winning and oscar. But including the SAG to the acting models and DGA to the director model, these relationships quickly fail or weaken. And even Best Picture seems to follow this patern, despite that the DGAs do not award on the merit of Best Picture.

The supporting actor/actresses seem to be the dark horses here, they are much less affected by the guild awards than their leading counterparts.

End result

We now have several models, both with and without the inclusion of the guild awards. While the prevalence of the globes and baftas as significant predictors is a good sign, there is still something to be desired.

- We have not been able to find any signifcant characteristics of the nominees (age, ethnicity, past wins...)
- We have not been able to find any significant characteristics of the films (rating, release date, adapted work)

All our variables come from the results of previous film awards. For the purpose of prediction, this is satisfactory. But for the purpose of description, these models are fairly bland. Given the drastic bivariate relationships seen in the EDA, we expected more nominee-specific effects to emerge.

Section 4

Cross-validation and Diagnostics

Post-estimation Diagnostics

- Check if the model fits the data (Deviance and Chi-square goodness-of-fit)
- Check for multicollinearity (Variance inflation Factors)
- Check model specification
- Check for linearity between covariates and logit

Cross-Validation

- Bootstrap
- K-fold cross-validation
- Historical performance

Section 5

Prediction

2014 Predictions

With a working model, our logistic regression model appears as follows:

$$\text{logit}(p) = \ln(\frac{p}{1-p}) = \alpha + \beta X$$

We will then transform this, and fit the data for the 2014 Oscar nominees to find predicted probabilities for this year's nominees.

$$\widehat{p} = \frac{e^{\widehat{\alpha}+\widehat{\beta}X}}{1+e^{\widehat{\alpha}+\widehat{\beta}X}}$$

2014 data

For now, we will not use the SAG Acting Model for prediction as I am not entirely comfortable making predictions from such a small sample size. Though, we have no qualms using the DGA Models for the Picture and Acting Race.

★denotes the 2014 Screen Actor's Guild Winner

★ denotes the 2014 BAFTA winner

★ denotes the 2014 Golden Globes winner

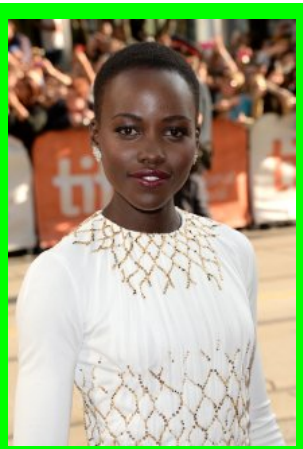
★ denotes the 2014 Director's Guild Award Winner

★ denotes the 2014 Critic's Choice award winner, though it is not modelled or used for prediction

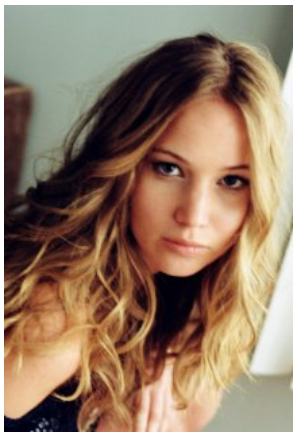
Best Actress in a Supporting Role

Sally Hawkins
P=.13±.03

Julia Roberts
P=.06±.02



Lupita Nyong'o
P=.43±.10



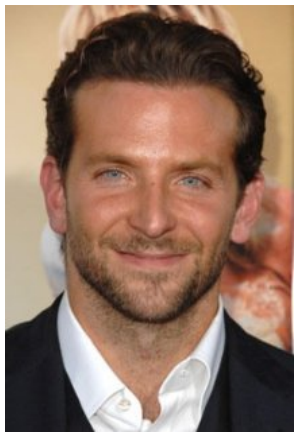
Jennifer Lawrence
P=.42±.08

June Squibb
P=.13±.03

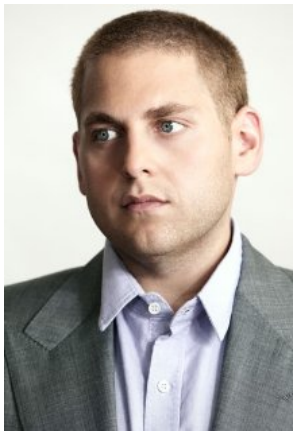
Prediction: Lupita Nyong'o wins for 12 Years a Slave

Best Actor in a Supporting Role

Barkhad Abdi
P=.11±.03
★



Bradley Cooper
P=.11±.03



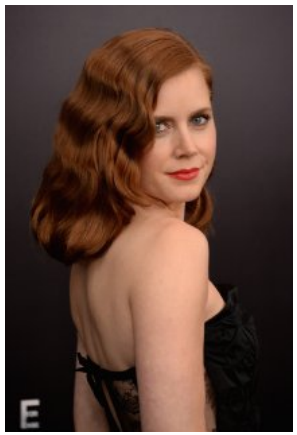
Jonah Hill
P=.11±.03

Michael Fassbender
P=.11±.03

Jared Leto
P=.68±.08
★★★

Prediction: Jared Leto wins for Dallas Buyer's Club

Best Actress in a Leading Role



Amy Adams
P=.16±.05

Cate Blanchett
P=.62±.13

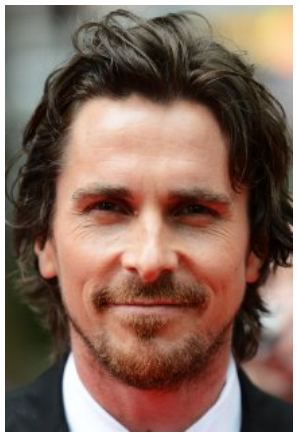
Sandra Bullock
P=.16±.05

Judi Dench
P=.15±.05

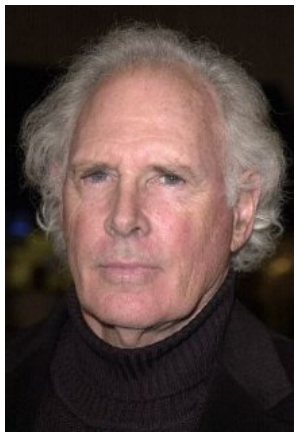
Meryl Streep
P=.06±.02

Prediction: Cate Blanchett wins for Blue Jasmine

Best Actor in a Leading Role



Christian Bale
P=.03±.03



Bruce Dern
P=.34±.03

Leonardo Dicaprio
P=.12±.03

Chiwetel Ejiofor
P=.36±.10

Matthew McConaughey
P=.60±.09

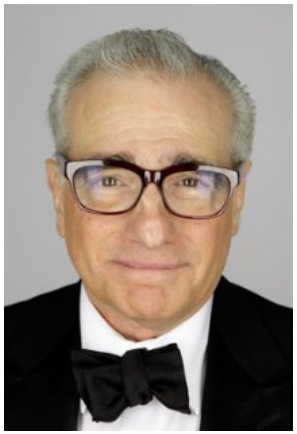
Prediction: Matthew McConaughey wins for Dallas Buyer's Club

Best Director

Alfonson Cuaron
P=.95±.03

Steve McQueen
P=.06±.02

David O' Russell
P=.06±.02



Martin Scorsese
P=.01±.01

Alexander Payne
P=.01±.01

Prediction: Alfonso Cuaron wins for Gravity

Best Picture

- American Hustle (P = 0.051±.02)
- Captain Phillips (p = 0.051±.02)
- Dallas Buyers Club (P = 0.051±.02)
- Gravity (P = 0.85±.07) ★
- Her (P = 0.00±.00)
- Nebraska (P = 0.00±.00)
- Philomena (P = 0.00±.00)
- 12 Years a Slave (P = 0.24±.11)★★★
- The Wolf of Wallstreet (P = 0.00±.00)

Where to go from here?

I will get more information once the 2014 Academy Awards have finished. However, some shortcomings and improvements are already evident. First, I must run rigorous diagnostics and validation methods on my models. I am very suspicious that my models are overfit.
Second, further analysis is necessary on the anomaly of my regular Acting and Directing Models, and the SAG-enhanced and DGA-enhanced models. The SAG and DGA variables are neither interacting nor multicollinear based on VIF and inclusion of interaction terms.