Are Best Actor Awardees More Likely to Be White?

Final Term Project - Data Analysis 2

Kata Süle

3rd January 2021

Abstract

The aim of this analysis is to find out whether Best Actor Awardees at the Academy Awards are more likely to be white than of other colours. To investigate this question I use a data set on Best Actor Awardees between 1929 and 2014 with race being the dependent variable and explanatory variables such as place of birth, sexual orientation, elapsed years since the award ceremony and age at award ceremony. By using these variables I estimate several probability models including linear probability models, a logit model and a probit model and come to the conclusion that neither of the models have significant parameters therefore they are not suitable for prediction purposes. Lastly, I check if my results have external validity for which I use samples of the Best Supporting Actor, Best Actress and Best Director categories and find that none of the parameters are significant in any of the models. Therefore I conclude that the insignificant results are probably not due to the small sample size but to inadequate explanatory variables.

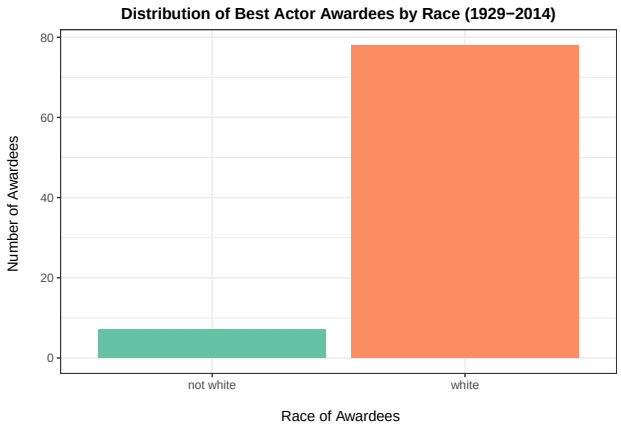
Introduction

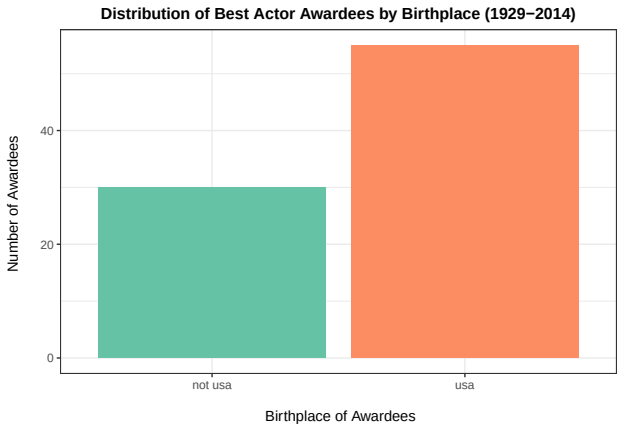
Discrimination has always been present in many parts of life. Probably discrimination against gender and race are the two types that make headlines most often and not so surprisingly also in connection with the Academy Awards. In this analysis I only focused on race discrimination because I was curious to see whether the data proves the statement that Best Actor Awardees at the Academy Awards are more likely to be white.

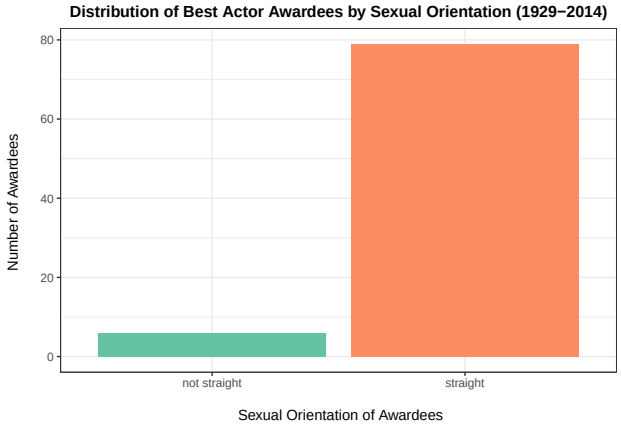
Data

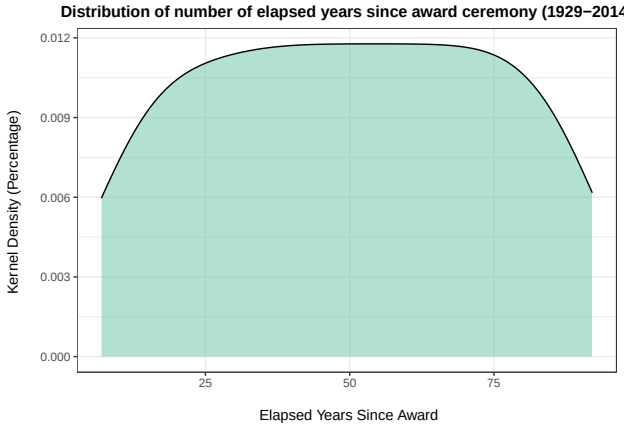
The sample that I used for the analysis contains 85 observations on Best Actor Awardees between 1929 and 2014. It is important to note that the sample does not have complete coverage since it does not contain data on awardees between 2015 and 2020. Furthermore, it also has selection bias since there is a strong filter on who gets to be nominated for the Best Actor Award. The dependent variable - *race* - is binary and its two possible outcomes are white and not white. There are four explanatory variables in total: *place of birth* which is binary and its two outcomes are USA and not USA, *sexual orientation* which is also binary and its two outcomes are straight and not straight, *elapsed years since awardee got the award* which is quantitative and measured on a ratio scale and lastly *age of awardee on receiving the award* which is also a quantitative, ratio variable. Out of the four explanatory variables *sexual orientation* is likely to have lower reliability since people are not always open about this topic which also means that the variable might suffer from measurement error. However, since my assumption was that it can be an important variable I decided to keep it in the sample.

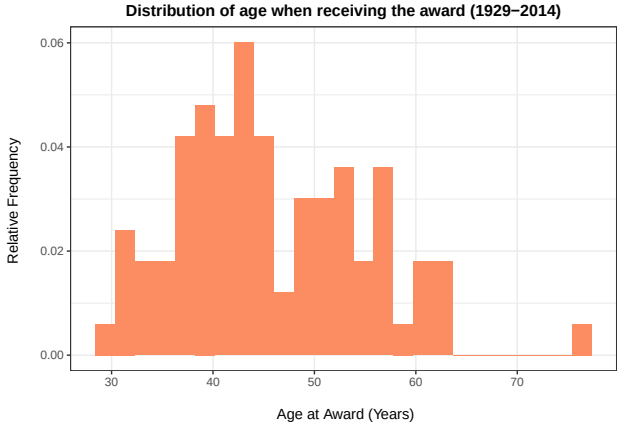
After creating the workfile I checked the distributions of the variables by using bar charts, histograms and kernel density plots which can be seen below. Based on the bar chart of the *race* variable we can say that the majority of actors are white, in addition by looking at the distribution of the *birthplace* variable we can conclude that most of them were born in the US. As for the *sexual orientation* variable the majority of actors are straight. If we look at the density plot of the *years since award* variable we can see that it approximates a uniform distribution. This is because except for 1930 there was only one awardee in every year. Lastly, the *age at award* variable has one extreme value, however since it is not an error and it is useful to have variation in the explanatory variables I decided to keep this observation. Otherwise, the distribution is a little skewed but since the ln transformation did not bring it closer to normal I opted for no transformation.











Besides the plots I also checked the descriptive statistics of the variables which can be seen in the table below. The means of the *race*, *birthplace* and *sexual orientation* variables show that the majority of actors are white, were born in the US and straight. We can see that the first award was given out 92 years ago and that actors are usually between their late thirties and early fifties when receiving the award.

Table 1: Descriptive statistics of variables

Variable	Mean	Median	Std	Min	Max	N
Race	0.92	1	0.28	0	1	85
Birthplace	0.65	1	0.48	0	1	85
Sexual Orientation	0.93	1	0.26	0	1	85
Years Since Award	49.87	50	24.98	7	92	85
Age At Award	45.87	44	9.03	30	77	85

Model

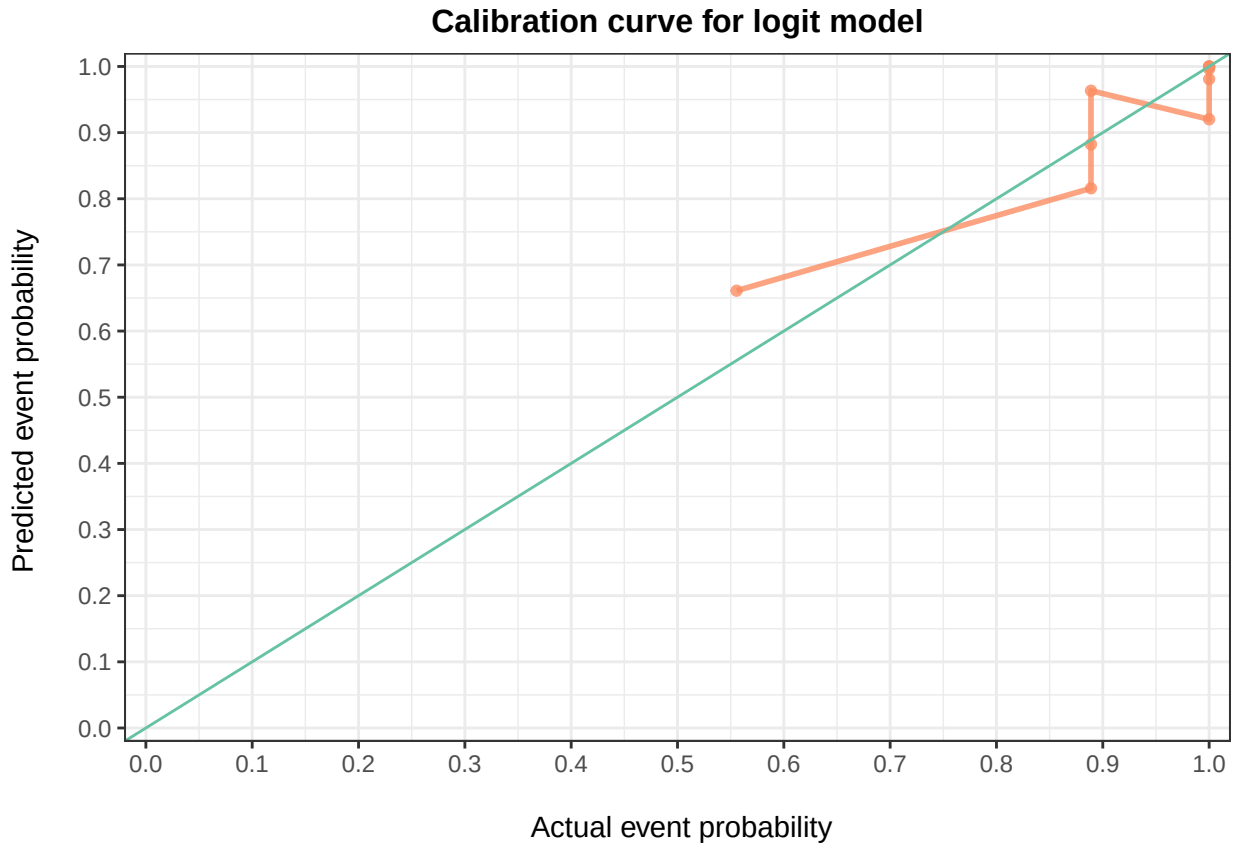
To find the answer to my research question first I inspected the pattern of association between the dependent variable and the quantitative explanatory variables (see Appendix). Based on the relationship check I decided to include the *years since award* variable in a linear way, whereas the *age at award* variable was included with a piecewise linear spline at age 47. Furthermore, I created an interaction between the *birthplace* and the *years since award* variables since I assumed that if an actor got the award a long time ago and was even born in the US then that would increase the probability of him being white.

In total I estimated five different models, three linear probability models, one logit and one probit model. After running the models I had to conclude that none of the parameters proved to be significant in my best three models (see Appendix). This was probably due to not having enough variation in my qualitative explanatory variables as well as the size of the sample.

According to the log likelihood value the logit model was slightly better than the probit model, therefore I only show the formula - where lambda indicates the link function - for this model without the parameter estimates since those were not significant.

$$y^p = \Lambda(\beta_0 + \beta_1 birthplace_i + \beta_2 sexual.orientation_i + \beta_3 years.since.award_i + \beta_4 age.at.award(<47)_i + \beta_5 age.at.award(>47)_i + \beta_6 birthplace * years.since.award_i)$$

To show that the logit model is not suitable for prediction I created a calibration curve which displays the goodness of fit of a model. The chart can be seen below. Since the orange line which indicates the predicted probabilities does not overlap with the green line showing the actual probabilities we can conclude that the model is not well calibrated at all.



Since it could have happened that I did not add some important variables that could have made the parameter estimates better I decided to do a robustness check for the logit model. To do this I estimated two additional logit models: one without the interaction variable and one with two knots for the *age at award* variable which was supported by the lowess graph between the *race* and the *age at award* variables. The comparison table of these two models and the original logit model can be seen below. Since all of the parameters are non-significant for the new models as well I concluded that the model is robust.

Table 2: Robustness check with two additional logit models

	Logit Orig.	Logit with Two Knots	Logit without Interaction
Constant	25.17	29.14	21.11
	(2405.28)	(17501.22)	(2493.24)
Birthplace	-2.88	-2.88	-0.02
	(2.18)	(2.18)	(0.93)
Sexual Orientation	-16.88	-20.85	-15.67
	(2405.28)	(17501.21)	(2493.24)
Years Since Award	-0.02	-0.02	0.02
	(0.03)	(0.03)	(0.02)
Age At Award (< 47)	-0.13		-0.11
	(0.11)		(0.10)
Age At Award (> 47)	0.92		0.82
	(0.78)		(0.68)
Birthplace * Years Since Award	0.07	0.07	
	(0.04)	(0.04)	
Age At Award (< 47)'		-0.13	
		(0.11)	
Age At Award (> 47 & < 56)		0.92	
		(0.79)	
Age At Award (> 56)		8.98	
		(2937.94)	
Num.Obs.	85	85	85
BIC	66.9	71.3	65.3
Log.Lik.	-17.883	-17.883	-19.304
Pseudo.R2	0.26	0.26	0.20

* p < 0.05, ** p < 0.01

```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```

External validity

Discrimination against race is said to be present in every award category at the Academy Awards therefore I ran the original logit model on three additional samples: Best Supporting Actor Awardees, Best Actress Awardees and Best Director Awardees. This way I could check if my findings had high or low external validity for other categories.

The comparison table below displays the results of the original logit model for Best Actor Awardees and the three additional ones mentioned above.

Table 3: External validity check on three new samples

	Actor	Supp. Actor	Actress	Director
Constant	25.173	11.963	41.368	497.895
	(2405.283)	(2244.205)	(39165.491)	(219844.018)
Birthplace	-2.884	-1.403	-23.262	29.203
	(2.177)	(1.682)	(31628.906)	(65089.174)
Sexual Orientation	-16.877	-14.585	-18.069	0.161
	(2405.276)	(2244.204)	(23098.658)	(72413.398)
Years Since Award	-0.020	0.003	0.004	0.285
	(0.033)	(0.021)	(511.288)	(0.288)
Age At Award (< 47)	-0.134	0.081	-0.003	-10.699
	(0.111)	(0.066)	(0.162)	(4416.505)
Age At Award (> 47)	0.923	0.070	5.094	0.109
	(0.784)	(0.066)	(1954.813)	(0.237)
Birthplace * Years Since Award	0.071	0.071	0.131	-0.297
	(0.044)	(0.046)	(511.288)	(1338.196)
Num.Obs.	85	73	85	86
BIC	66.9	71.7	38.2	35.5
Pseudo.R2	0.260	0.235	0.350	0.773

* $p < 0.05$, ** $p < 0.01$

fitting null model for pseudo-r2 fitting null model for pseudo-r2 fitting null model for pseudo-r2 fitting null model for pseudo-r2

Based on the comparison table we can see that the coefficient estimates are similar to the original ones except for the sample of directors where they tend to be much higher. However, what is important is that none of the parameters are significant therefore I can conclude that the results of the original model do not have external validity for the Best Supporting Actor, Best Actress and Best Director categories. This suggest that the insignificant results of the original model are probably due to the used variables and not the sample size.

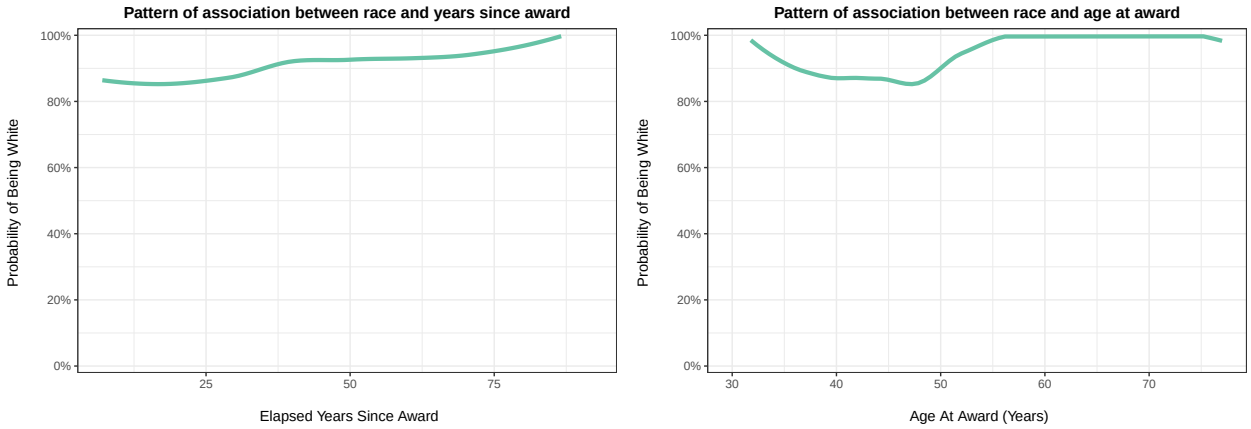
Summary

In this analysis I investigated whether Best Actor Awardees at the Academy Awards are more likely to be white with the help of probability models. I estimated linear probability models, a logit and a probit model with race being the dependent variable and four explanatory variables. Based on the results I had to conclude that none of the parameters of my models were significant therefore they are not suitable for prediction purposes. Furthermore, by checking the external validity of the model I saw that the insignificant results were probably caused by inadequate variables and not small sample size. Which means that the place of birth, the sexual orientation, the elapsed years since receiving the award and the age at award are not suitable to predict the race of an awardee.

Appendix

Inspection of pattern of association

To be able to decide about which functional form would be the best for the quantitative explanatory variables in the estimation I created a graph with lowess for each of them. Based on these I concluded that there is a linear relationship between the *race* and the *years since award* variables, whereas for the *age at award* there seems to be a knot at 47 years and one at 56 years.



Models and their characteristics

Since my models did not have significant parameters most of this section aims to show what kind of insights we can get from probability models given that they are significant.

Estimated models The model comparison table below shows the coefficients and the marginal effects - if applicable - of the best three models. None of the coefficient are significant, however since the logit model is a bit better than the probit according to the log likelihood value I decided to use this model to show how we can evaluate probability models in the upcoming sections.

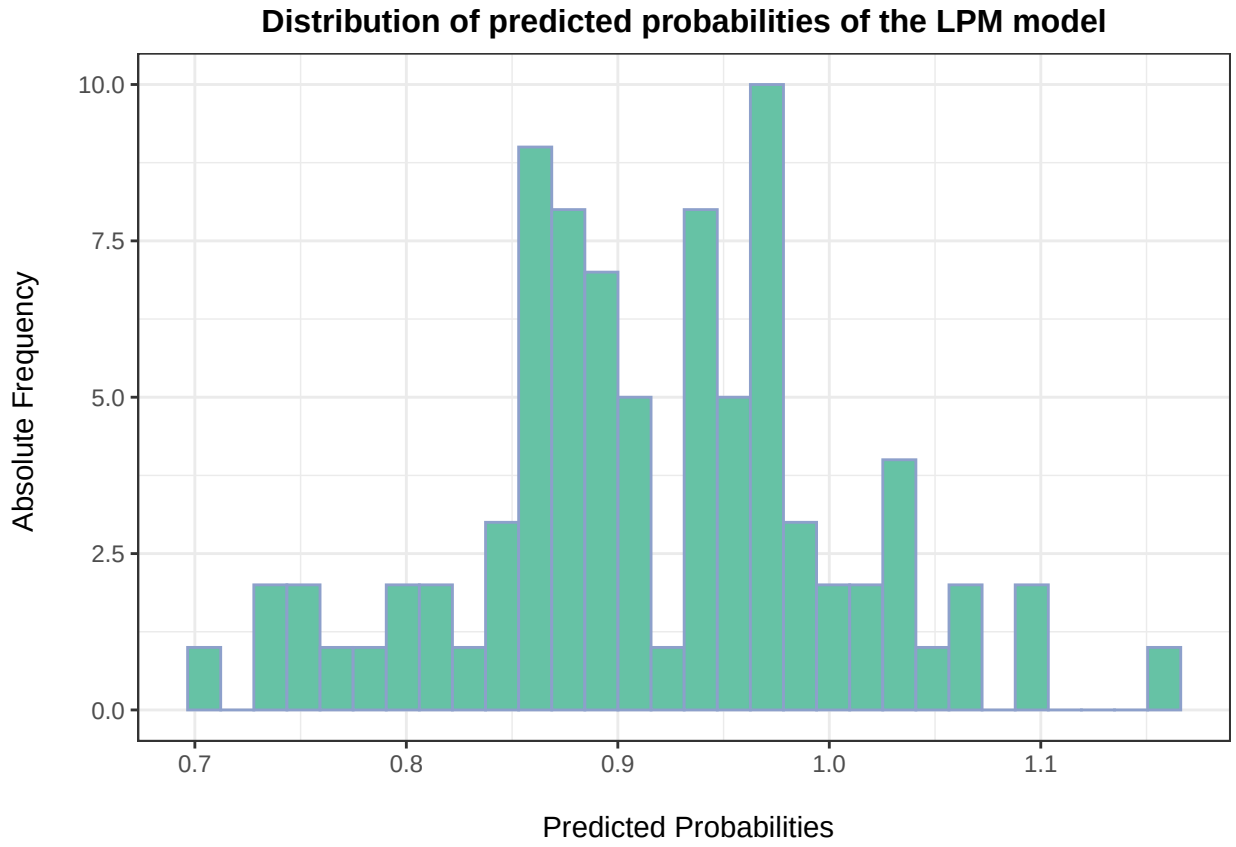
fitting null model for pseudo-r2 fitting null model for pseudo-r2

Table 4: Results of estimated models

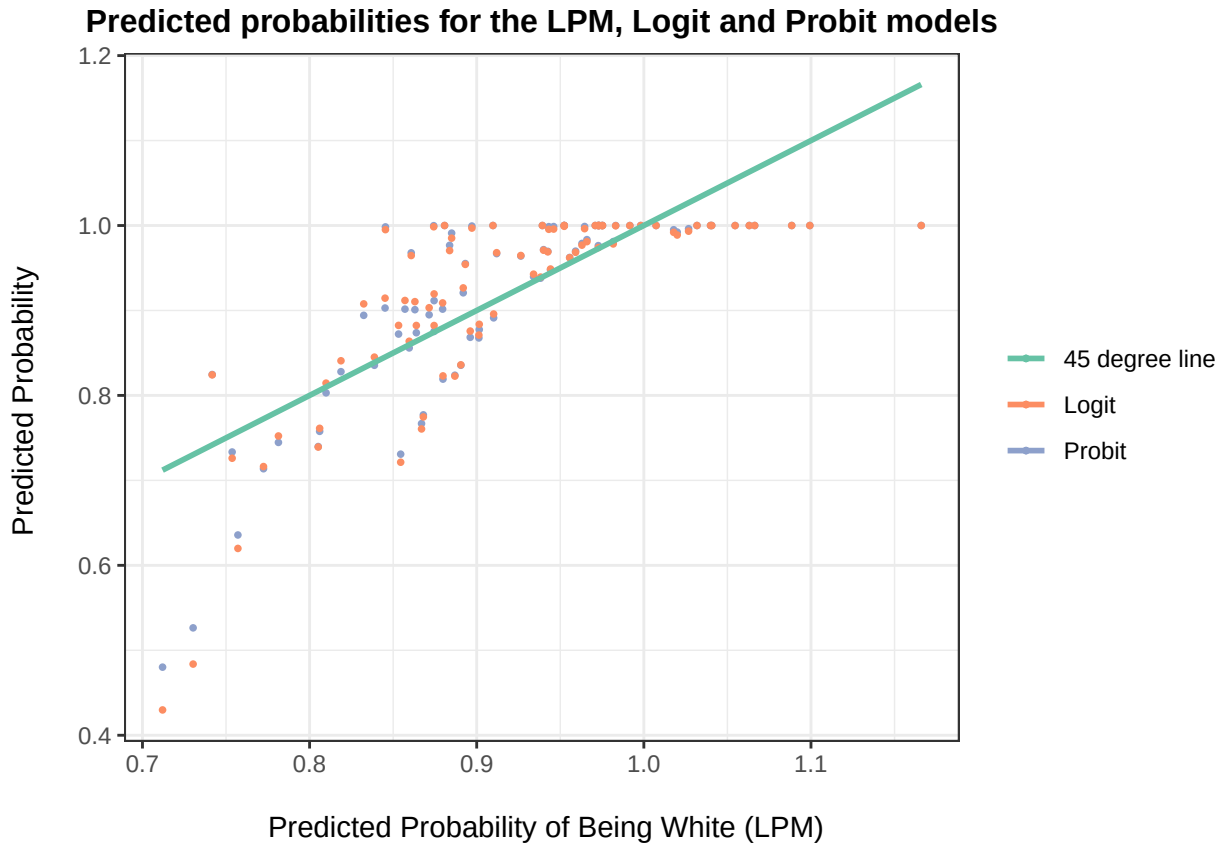
	LPM + PLS	LPM + PLS + Int.	Logit	Logit Mfx	Probit	Probit Mfx
Constant	1.07**	1.28**	25.17		9.40	
	(0.33)	(0.35)	(2405.28)		(554.28)	
Birthplace	-0.01	-0.23	-2.88	-0.15	-1.53	-0.15
	(0.06)	(0.14)	(2.18)	(0.08)	(1.16)	(0.08)
Sexual Orientation	-0.06	-0.11	-16.88	-0.09**	-5.07	-0.09**
	(0.12)	(0.12)	(2405.28)	(0.03)	(554.27)	(0.03)
Years Since Award	0.00	-0.00	-0.02	-0.00	-0.01	-0.00
	(0.00)	(0.00)	(0.03)	(0.00)	(0.02)	(0.00)
Age At Award (< 47)	-0.00	-0.01	-0.13	-0.01	-0.07	-0.01
	(0.01)	(0.01)	(0.11)	(0.01)	(0.06)	(0.01)
Age At Award (> 47)	0.01	0.01	0.92	0.06	0.49	0.06
	(0.01)	(0.01)	(0.78)	(0.03)	(0.42)	(0.03)
Birthplace * Years Since Award		0.00	0.07	0.00	0.04	0.00
		(0.00)	(0.04)	(0.00)	(0.02)	(0.00)
Num.Obs.	85	85	85	85	85	85
R2	0.065	0.101				
BIC	47.1	48.2	66.9	66.9	66.9	66.9
Log.Lik.	-7.986	-6.323	-17.883	-17.883	-17.904	-17.904
Pseudo.R2			0.26		0.26	

* $p < 0.05$, ** $p < 0.01$

Predicted probabilities The histogram below aims to show that the predicted probabilities of the linear probability model can be higher than 1 which makes it unsuitable for prediction purposes. However, it is visible that the probability of an actor being white is relatively high throughout the sample.



This chart shows how the predicted probabilities of the three models are related to each other. We can see that the predictions of the logit and probit models are very close to each other, however they both differ a lot from the linear probability model. The chart also shows that the probit and the logit models do not predict probabilities above 1, unlike the linear probability model.



These two tables show the characteristics of actors who belong to the bottom and top 10% based on the predicted probabilities by the logit model. We can see that actors who are in the bottom 10% - meaning that their probability of being white is the lowest - were born in the US, are straight, received the award 19 years ago on average and were generally 44 years old when receiving it. As for the top 10% in comparison, 88% of actors were born in the US, 75% of them is straight, received the award much longer ago, 73 years on average, and were also older on average when receiving it.

\begin{table}[!h]

\caption{Characteristics of bottom 10% (Logit)}

Statistics	Birthplace	Sexual Orientation	Years Since Award	Age At Award
mean	0.78	1	29.11	43.56
median	1.00	1	17.00	45.00
sd	0.44	0	25.04	3.36

\end{table}

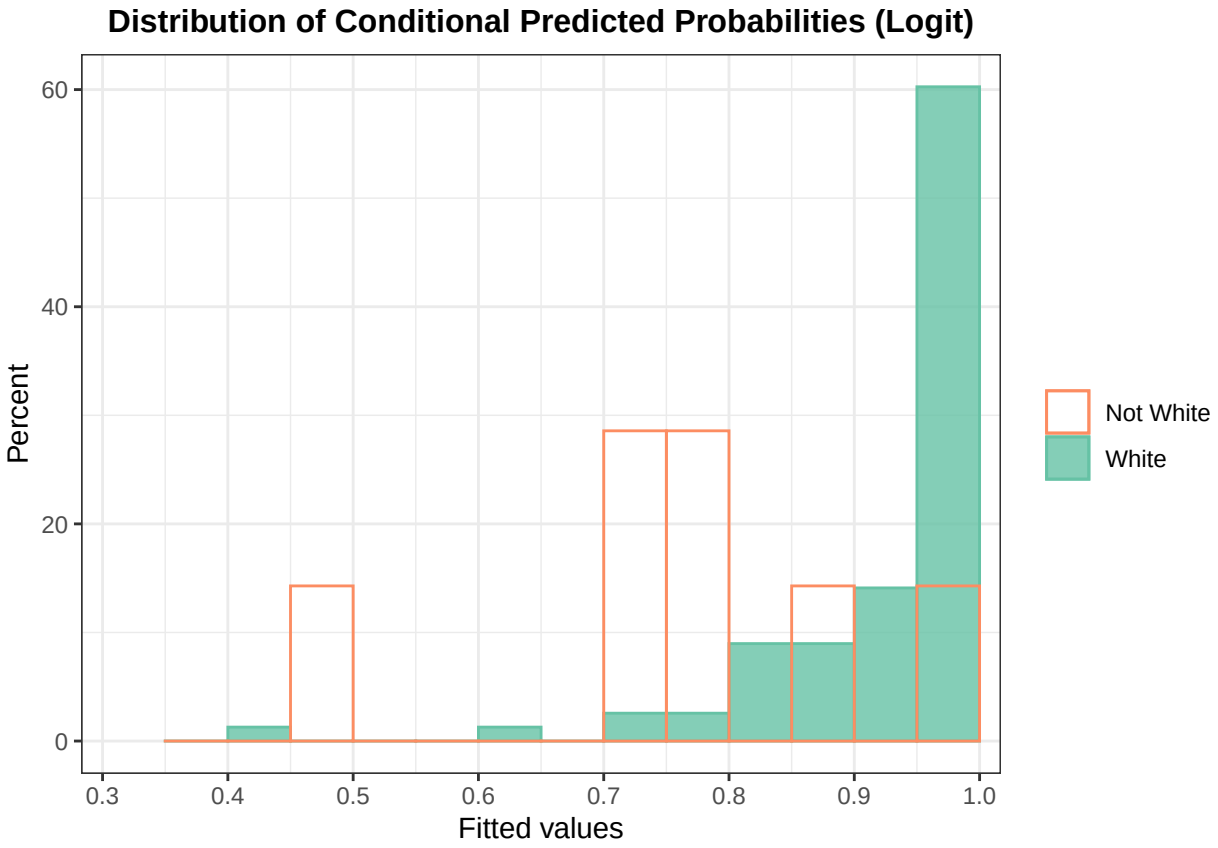
\begin{table}[!h]

\caption{Characteristics of top 10% (Logit)}

Statistics	Birthplace	Sexual Orientation	Years Since Award	Age At Award
mean	0.50	0.25	64.75	50.88
median	0.50	0.00	64.50	47.00
sd	0.53	0.46	18.62	14.22

\end{table}

Conditional predicted probabilities The purpose of this histogram is to show the distribution of probabilities predicted by the logit model conditional on the race variable. We can see that the predictions for the two groups overlap and that the model sometimes predicts low probabilities for actors who are actually white and that it also predicts high probabilities for actors who are not white. This proves that the model is not really suitable for prediction.



The two tables below show the conditional probabilities predicted by the three models. For white actors we can see that the biggest difference between the models are in their minimums. The linear probability model has a minimum of 0.71, while the logit and the probit models have theirs at 0.43 and 0.48. As for non-white actors the pattern is very similar with the minimums differing the most, however in this case the standard deviation is also higher for the logit and the probit models.

Table 5: Summary statistics of predicted probabilities if race is white

Statistics	LPM	Logit	Probit
mean	0.93	0.93	0.93
median	0.94	0.97	0.98
min	0.71	0.43	0.48
max	1.17	1.00	1.00
sd	0.08	0.10	0.10

Table 6: Summary statistics of predicted probabilities if race is not white

Statistics	LPM	Logit	Probit
mean	0.83	0.76	0.77
median	0.81	0.75	0.74
min	0.73	0.48	0.53
max	0.93	0.96	0.96
sd	0.08	0.15	0.14