Predicting the Best Director at the Oscars with Bayesian Logistic Regression

Glen Morgenstern and Jason Gerber

4/24/2021

Introduction

Every year, billions of people around the globe watch movies, and many tune in to watch the Academy Awards. This awards show is heavily publicized, promoted, and talked about every year, where millions of fans watch the ceremony and tens of millions read about it later. Receiving an Oscar is the highest honor that one can procure in the business of movie-making, and since movies are such a large industry and many of our country's most influential celebrities are movie stars, the Oscars is one of the biggest events there is each year in the United States. All this attention inevitably leads to general curiosity and predictions about who will take home the most prized awards in all of filmmaking, which is where the reason for this project arises.

The goal of this project is to develop a model which will predict with as much accuracy as possible the winner of the "Best Director" category at the Oscars. Being able to predict Oscar winners is an analysis which many across the globe will have an interest in, whether they be data-lovers, typical cinephiles, or even gamblers who feel that they need an edge.

To achieve our goal, we will take several steps. First, we will perform exploratory data analysis on the data set we have in order to discover which variables have the potential to be explanatory variables for the model. Second, we will establish a clear prior for the model. Third, we will determine which explanatory variables should be used for the model by combining them with the prior to form a posterior, by creating several Bayesian logistical regression models and testing until we reach a model with optimal fit. Lastly, we will interpret the model coefficients and discuss its applications.

Data

The data set that we are using for this analysis contains information about all Oscar nominations from 1928-2006. The data have been taken from the personal website (http://www.stat.columbia.edu/~gelman/arm/examples/academy.awards/) of Iain Pardoe, an online educator in mathematics and statistics, in the form of a csv, which we analyzed in R.

In our model, we will attempt to use all relevant variables that may be significant predictors of whether a director wins Best Director or not. The data set given has 62 variables, of which one is an indicator variable specifying if the nominee is up for Best Director, and one of which is an indicator variable specifying if the nominee won their award or not. We will create our data set using only the instances where DD is equal to 1, as we are only focusing on predicting winners among nominated directors.

To decide which variables to include in our model, we performed EDA on many variables, but since we have limited space, we are only discussing the variables which might be relevant to our model. In other words, we will not display EDA on variables like AD (Whether the Assistant Director was nominated) because it simply did not have any relevance or correlation with the director winning Best Director.

The variables that we performed EDA on are described and displayed below:

- Nom: The total number of Oscar nominations that the director's movie received.
- Pic: Director's movie was also nominated for best picture (indicator).
- Scr: Director's movie was also nominated for best screenplay. We are unsure why there would be any reason why Scr is equal to 2, as the same movie cannot be nominated twice for best screenplay. We believe that is an issue in the data, though miniscule.
- Cin: Director's movie was also nominated for best cinematography (indicator).
- Edi: Director's movie was also nominated for best editing (indicator).
- Eff: A variable that specifies if the director's movie was nominated for sound editing, visual effects, or both. For both, Eff = 2, for one and not the other, * Eff = 1, and for neither, Eff = 0.
- Gd: Director's movie won a Golden Globe for Directing in the same award season (indicator).
- DGA: Director's Guild of America win (indicator).
- Year: The year that the movie was nominated for an Oscar.

Exploratory Data Analysis

Plot 1: Total Oscar Nominations
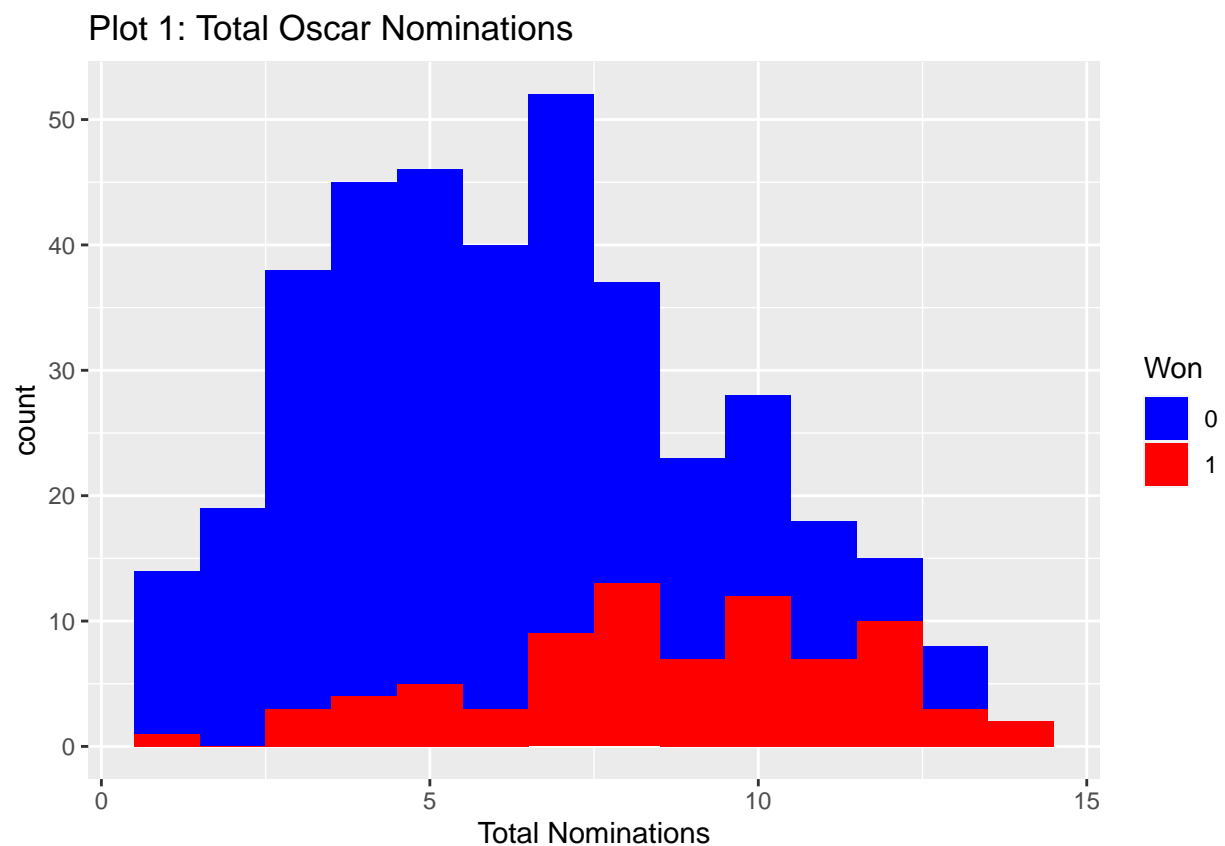
The above analysis makes it appear that the more nominations a movie has, the more likely the director is to win an Oscar. This is enough correlation to warrant adding it to our initial model.

Plot 2: Various other EDA
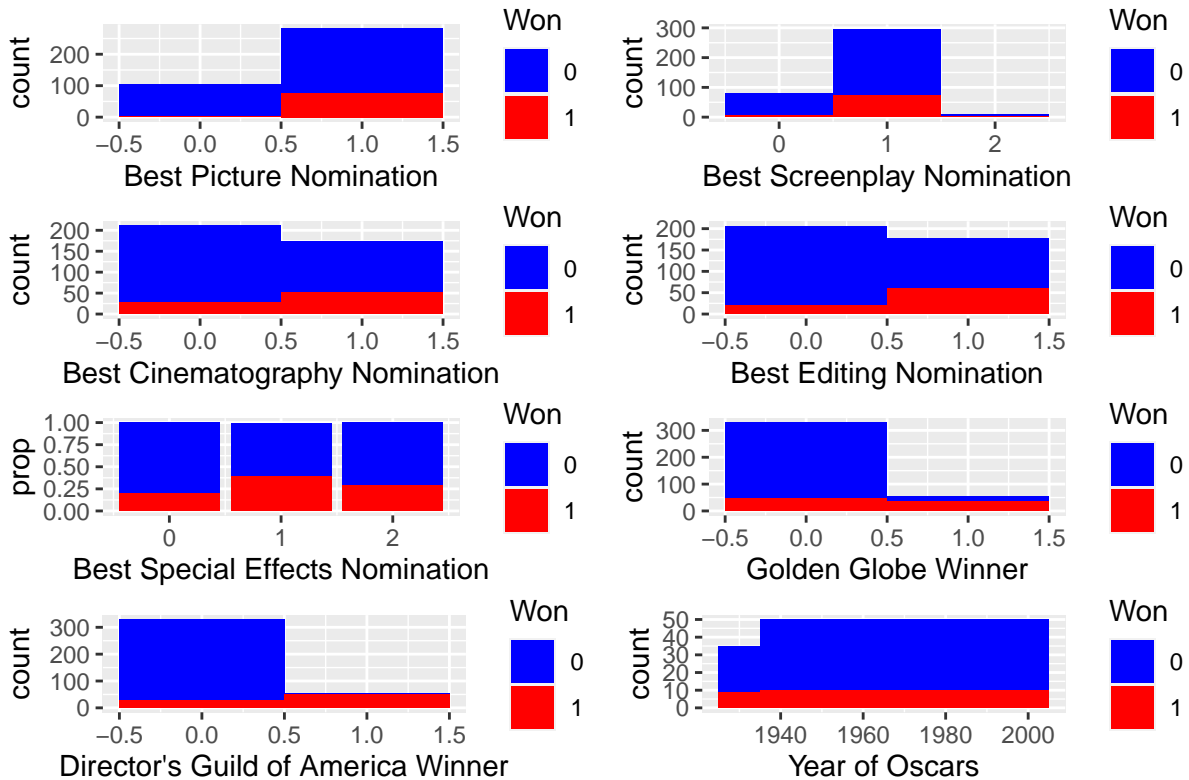
All of the above variables appear to show some correlation in a difference between an increase in them and an increase in the proportion of wins (except for year, which has a greater proportion of wins in the first decade).

The proportion of Best Director winners given a best picture nomination is much higher than without a best picture nomination, and similarly, the proportion of Best Director winners also increases as Screenplay, Cinematography, Editing, Golden Globe winner, and Director's Guild of America winner go from 0 to 1. Even Special Effects appears to have a larger proportion of Best Director winners given 1 or 2 special effects nominations. Lastly, year clearly has a higher proportion of wins in the first decade of the Oscars.

We have chosen all of these variables as our possible predictors: DGA, Year, Gd, Edi, Scr, Cin, Pic, Nom, and Eff.

Model and parameters

We set a relatively weakly informative prior model of Normal (0, 10) for model coefficients. We do this because we have rather little practical prior information on the effects of each regressor. We also wanted to leave open the possibility that a coefficient could have a negative effect on Oscar chances. In addition, Pardoe found that using more informative priors based on results from each previous year produced equivalent, but not better results. (Pardoe)

We assumed independent priors for the purpose of this problem because while we do not know much about the Academy itself, we assumed that no film companies bribed members to vote for a certain movie in all categories. The intercept prior was set to Normal(0,1) because it led to lower values of looic.

Our sampling model used a logit link rather than a probit link for ease of interpretation. Below is the likelihood for a single observation under this model:

$$\binom{n}{y}\left(\frac{e^{\eta}}{1+e^{\eta}}\right)^{y}\left(\frac{1}{1+e^{\eta}}\right)^{n-y}$$

Figure 1: Likelihood

We used the stan_glm function in the **rstanarm** package to sample from the posterior. This function performs full Bayesian estimation via the Markov Chain Monte Carlo method. Below is our final model, with a graphic in Appendix I:

Table 1: Final Model

	2.5%	97.5%
(Intercept)	44.38	127.49
Pic	-0.08	2.76
Scr	-0.29	1.37
Edi	0.18	1.86
Cin	-0.40	1.16
DGA	4.12	6.68
Year	-0.07	-0.02

This can be compared to a model with total nominations and whether the candidate won Golden Globe for Best Director.

Table 2: Compared Model with More Variables

	2.5%	97.5%
(Intercept)	47.49	135.21
Nom	-0.42	0.08
Pic	0.06	3.19
Scr	-0.09	1.78
Cin	-0.30	1.52
Edi	0.20	2.33
Eff	-0.41	1.76
Gd	-0.52	1.73
DGA	4.22	7.01
Year	-0.07	-0.03

Results

Posterior Predictive Diagnostics

We assessed the strength of our model via its posterior preditive LOOCV and found a looic of 202.9. However, this accuracy rate is quite meaningless unless we have something to compare it to. We created a baseline model with only an intercept to compare to our final model. Its looic value was 735.1.

Loo2 (LOOCV for our final model) is much lower than the intercept-only model. And now we compare our final model with the model with the extra coefficients for total nominations and the Golden Globe. Its looic was 205.3, which is larger than our final model.

We computed posterior predictive probabilities of the linear predictor via the posterior_linpred() function provided in the **rstanarm** package. We calculate these posterior predictive probabilities in order to determine the classification accuracy of our model. This method gives us 91% accuracy.

4

We could also evaluate the classification accuracy of our model on unseen data. This can be done via a LOOCV approach. The results are similar to above, with about a 91% accuracy rate.

Convergence diagnostics

Appendix II shows some traceplots of the variables in the final model. The model clearly mixes rather well and appears stationary.

Interpreting Model Coefficients

If all explanatory variables are equal to zero, the log-odds are equal to 82.384. The intercept from the model with no predictor variables is equal to 82.384 as an estimation for the log-odds of winning Best Director.

Changing Pic from 0 to 1 and holding all else equal results in a change of 1.159 in the log-odds. This means that the film being nominated for best picture multiplies the odds of winning Best Director by 3.19 times what it would be if film were not nominated for best picture.

Increasing Scr from 0 to 1 and holding all else equal results in a change of 0.522 in the log-odds. This means that the film being nominated for best screenplay multiplies the odds of winning Best Director by 1.69 times what it would be if film were not nominated for best screenplay.

Changing Edi from 0 to 1 and holding all else equal results in a change of 1.024 in the log-odds. This means that the film being nominated for best editing multiplies the odds of winning Best Director by 2.78 times what it would be if film were not nominated for best editing.

Changing Cin from 0 to 1 and holding all else equal results in a change of 0.379 in the log-odds. This means that the film being nominated for best cinematography multiplies the odds of winning Best Director by 1.46 times what it would be if film was not nominated for best cinematography.

Changing DGA from 0 to 1 and holding all else equal results in a change of 5.281 in the log-odds. This means that the Director winning at the Director's Guild of America multiplies the odds of winning Best Director by 196.6 times what it would be if the director did not win at the Director's Guild of America.

Increasing Year by 1 and holding all else equal results in a change of -0.044 in the log-odds, so each year sees odds multiply by 0.96. However, this is only because from 1932 to 1935, there were just three nominees for Best Director in each of these years.

Conclusion

Our final model made use of the predictors Pic (best picture nomination), Scr (best screenplay nomination), Edi (best editing nomination), Cin (best cinematography nomination), DGA (winner of Director's Guild of America), and year (which was a nominee due to a lower amount of nominees in the first 4 years of the Oscars). One of the things that stood out was that having the best fit model (as defined by the lowest looic) did not mean having only variables that had a p-value less than 0.05, which is often a traditional mark of a good model with significant predictors.

Something else that we discovered was that the winner of the Golden Globe Best Director and the Director's Guild winner were highly multicollinear. Replacing DGA with Gd as a predictor in our model did not worsen the model greatly and allowed Gd to become significant (whereas it did not appear to be significant if DGA was also in the model). This is why we only included the Director's Guild winner variable in our final model. This variable was a very strong predictor of Oscar success, as the odds of a DGA winner taking the Oscar was 200 times that of a non-winner. Even so, being the winner of both the DGA and the Golden Globe is

not quite the bellwether that some have thought it to be, with less than 50% of Oscar-winning directors also winning a Golden Globe in the same year (Reid). Other characteristics of the director's movie are also very strong predictors of winning the award.

In addition, our model turned out to be highly accurate at predicting the winner of Best Director (within the data given), with a 91% success rate. This comes close to Iain Pardoe's 93% figure, and is only slightly less accurate. However, it is difficult to tell if this model would be very informative from a prediction standpoint. Data that informed this model range from 1928 to 2006, which leaves out more than a decade of rapid social and cultural changes that affect even the biggest stage in Hollywood. It is likely that this model would have changed since 2006, perhaps to accomodate values of equality and fairness that Hollywood has been forced to reckon with in recent years.
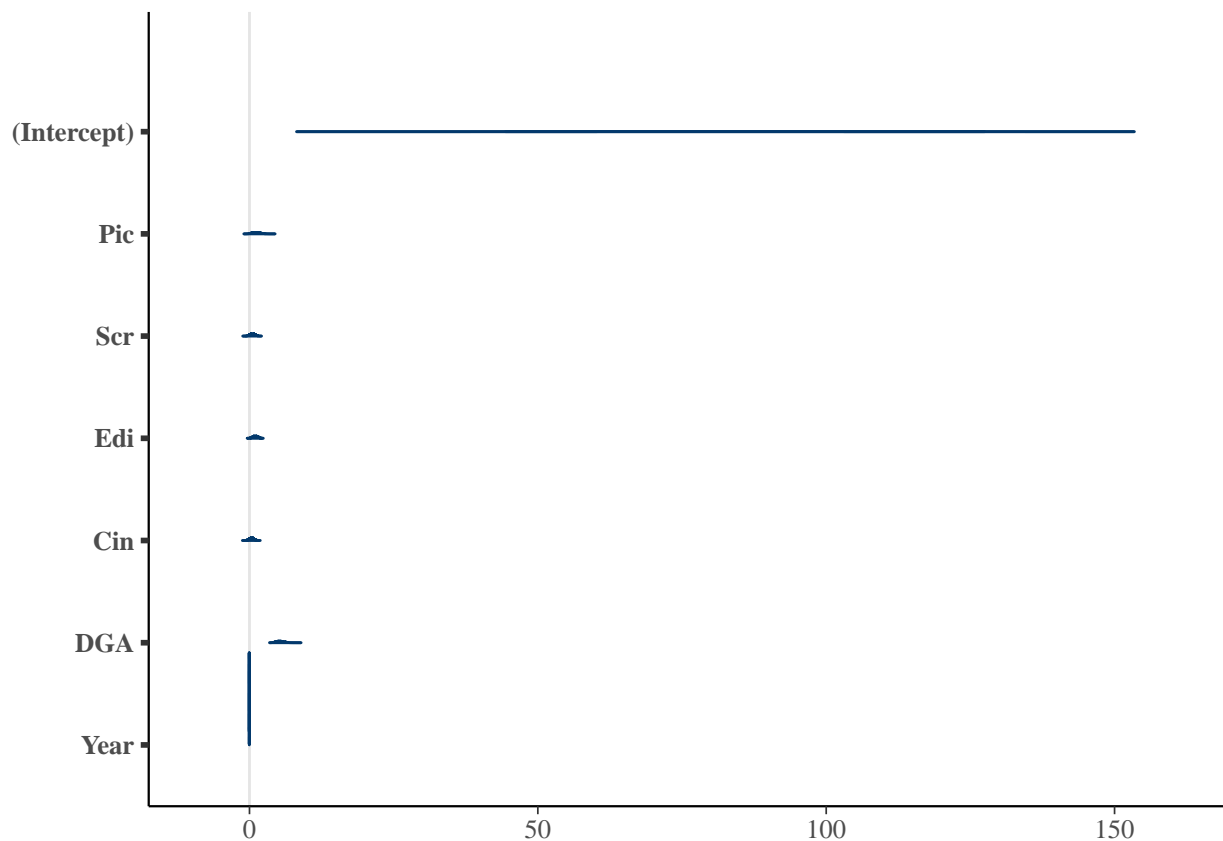
References

Pardoe, Ian. "Applying discrete choice models to predict Academy Award winners." Journal of the Royal Statistical Society, 171, Part 2, 2008, pp. 375-394. https://iainpardoe.com/refereed/08jrssa/oscars.pdf.

Reid, Joe. "How Likely Are Golden Globe Winners to Win Oscars Anyway?". Vulture. 3 Mar. 2021. https://www.vulture.com/2021/03/how-likely-are-golden-globe-winners-to-win-oscars-anyway.html .

Reproducible code: https://github.com/glenmorgenstern/morgenstern-gerber-oscars-bayesian

Appendix

Appendix I

Appendix II

Parameter: Pic

Parameter: Scr
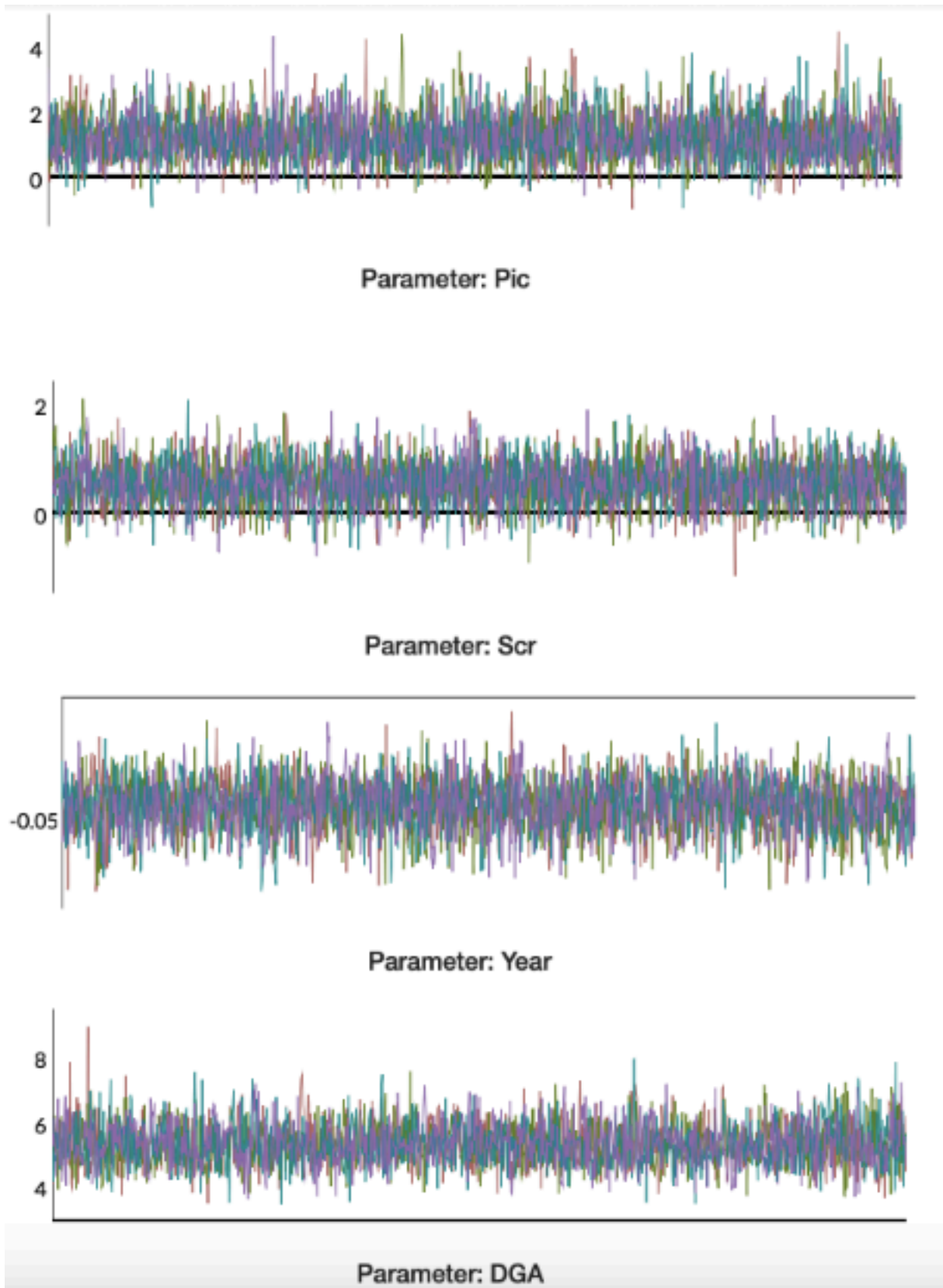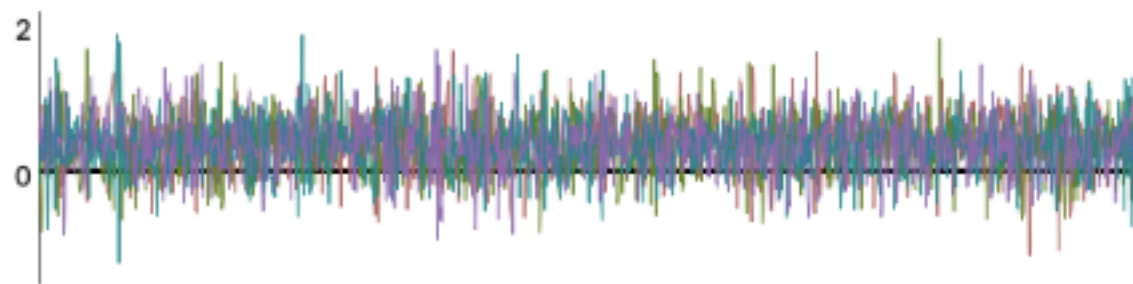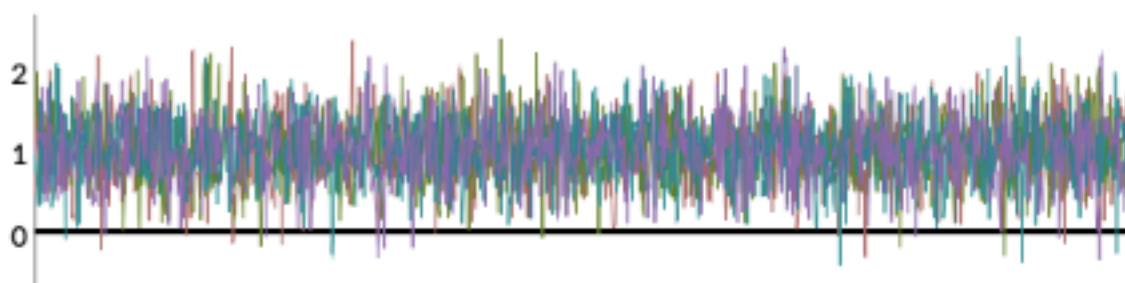
Parameter: Year

Parameter: DGA

Figure 2: Convergence Diagnostics

Parameter: Cin

Parameter: Edi

Figure 3: Convergence Diagnostics, cont.