The Pitchfork Formula

Robert Fry

1 Introduction

Anyone who is interested in modern indie rock knows that Pitchfork is unquestionably the top source for independent music discovery. "Entry level alts" flock to the website for music recommendations. Hipsters deny following it while, in reality, they religiously absorb every song that is approved on Pitchfork's "Playlist" (their recommended songs list). Pitchfork's "Best New Music" tag is a universal seal of popular approval; once an album appears on Pitchfork, it has already become mainstream by indie music standards. As an avid music fan, I was personally interested in determining whether there is any method to Pitchfork's rating scheme, and whether there are certain factors that reveal how well an album will be received. If there are, is there any way to use those indicators to predict the score that Pitchfork gives an album before the album is released and anticipate the levels of hype that the album receives? In order to test this, I created a multivariate regression that attempts to predict the Pitchfork score of an album using data on the Pitchfork website prior to the album's release. But perhaps the most relevant question that I hope to answer through this project is: "What should I listen to next?"

2 Data Collection

All of my data were collected from Pitchfork's website. I compiled a list of the 150 albums most recently reviewed by Pitchfork, excluding albums that I deemed were not applicable to my study (for example, reissues and albums released by bands with previous albums that were released before Pitchfork was created). After the process was done, I realized that artists with debut releases most often did not have enough recorded data on Pitchfork for them to be useful to my study,

---

This paper was written for Lauren Lax's Advanced Placement Statistics class in the spring of 2011.

and I excluded those, too (I did succeed, however, in using these data points to create a separate model to predict the scores of debut albums, although this model proved to be much less reliable). Because there were so few high scoring albums out of my first selection, I collected additional data from a list of the most recent high scoring albums. The final list contained 81 albums. For each album, I carefully recorded a number of characteristics about each album's respective artist. These factors are as follows: average album score, most recent album score (including EPs), least recent album score, highest album score, lowest album score, album number, number of tracks from the forthcoming album appearing on Pitchfork's "Playlist," number of tracks from the forthcoming album awarded "Best New Music," number of times the artist has previously appeared on a year-end list (eg: Best Albums of 2007), total number of tracks appearing on Pitchfork's "Forkcast."

I collected data from only the most recent albums since there have been noticeable shifts in Pitchfork's rating system, including the so-called "great inflation of 2k10," and the most recent albums would most accurately represent the recent rating trends. That said, I still neglected to perform a simple random sample in order to gain a representative sample of the album scores, which makes it possible that the data I collected represent an irregular period in Pitchfork ratings. Ideally, I would have defined a relatively recent period during which I believe ratings were relatively consistent and performed a simple random sample on albums from that period to determine which albums I incorporated into my sample. However, it was much more convenient to work chronologically, given that I could not anticipate time restraints, and that data collection took a particularly long time due to the number of explanatory variables I was collecting. In addition to affecting the randomness of my sample, the time restraints also prevented me from collecting a large number of observations. Although 81 albums is a considerable number of data points, more albums would have produced a more powerful regression.

The data table that includes all of the albums and their corresponding factors is too large to be presented in this section and is not crucial for the interpretation of the data, but can be viewed online at http://roundtable.menloschool.org.

3 Results

After compiling my data, I used Fathom to create a model to predict album ratings. According to the R^2 value of the multiple regression, approximately 56.42% of the variation in the rating an album receives from Pitchfork can be explained by the variation in the previously outlined explanatory variables. By performing a linear regression t-test on the coefficient of each of the explanatory variables, the statistical significance of each can be determined. The test statistics and P-values are listed along with the coefficient of each term. The initial model is displayed in Figure 1.

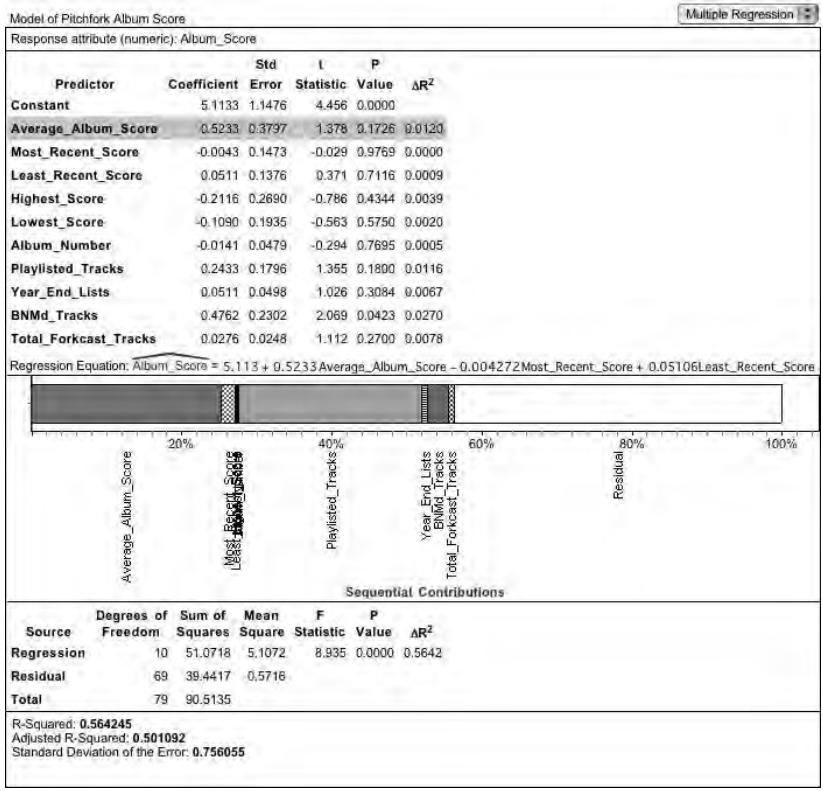


Figure 1: *Initial model*

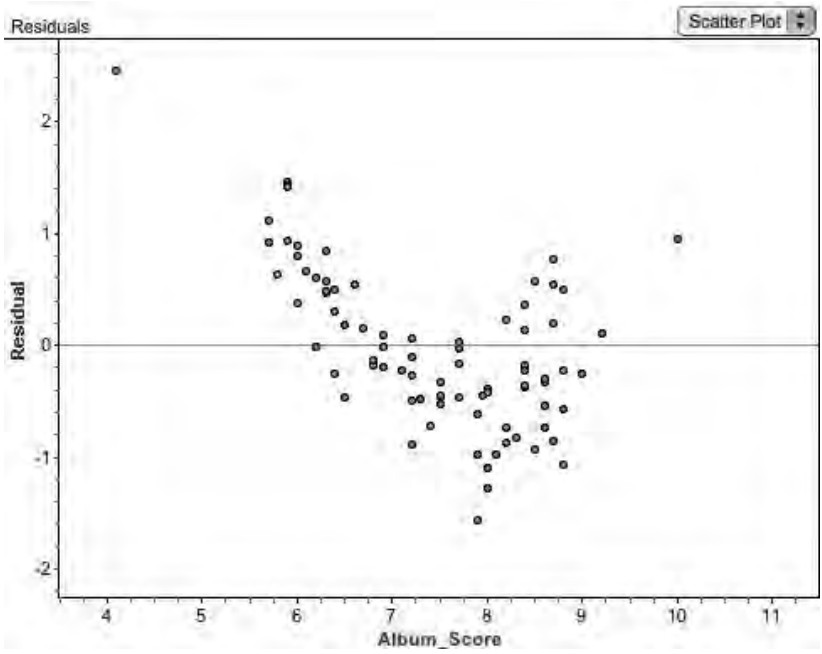
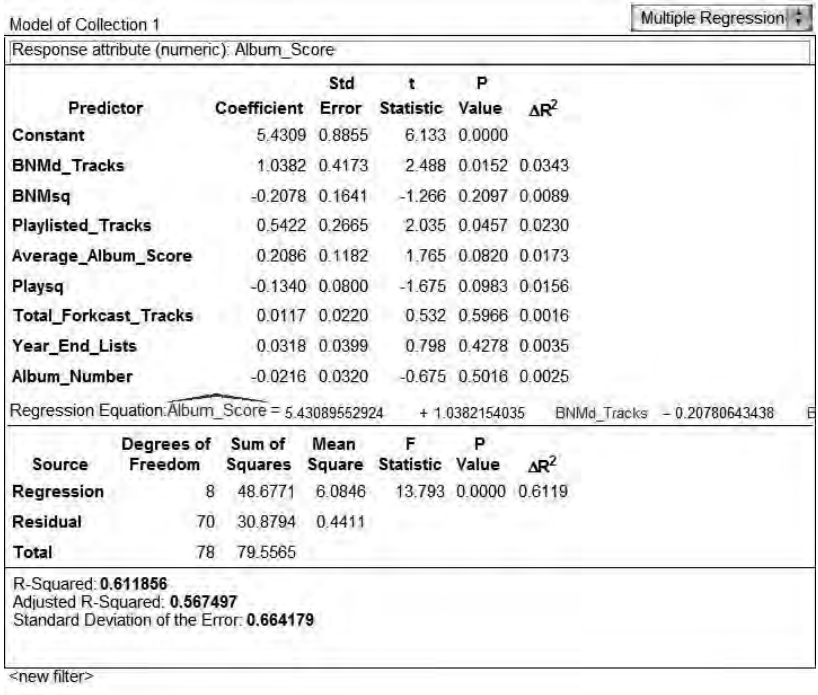


Figure 2: *Residual plot*

Unfortunately, as illustrated by the residual plot in Figure 2, there is a clear non-linear pattern to the residuals. Although the residuals seem approximately normally distributed, with no clusters and only slight heteroskedasticity, the errors are obviously not independently distributed, and therefore, the model does not meet the conditions for the inference test. The non-linearity of the residual plot could be attributed to the fact that the two most powerful predictors of album score are the number of Best New Music tracks and tracks on the Playlist. These, however, are rarely given out, so if they are, the model overvalues them in order to compensate. On the other hand, because the majority of albums score somewhere between 6 and 8, the baseline score that the model assumes is 5.1133. From there, there are no strong, independent factors that are negatively correlated, making it difficult for albums to be predicted to receive low scores.

Since the residuals appeared to follow a distinct pattern, I initially believed that a quadratic transformation of the data would be able to correct the model's misprediction. However, despite all attempts to modify predictors exponentially, no transformation could entirely eliminate the non-linear pattern. Still, by squaring the most significant predictors (Best New Music tracks and Playlist tracks), I found that this final model could achieve a slightly higher R^2 value ($R^2 = 0.6119$) and better account for the curve of the residuals.



Model of Collection 1 — Multiple Regression

Response attribute (numeric): Album_Score

Predictor	Coefficient	Std Error	t Statistic	P Value	ΔR^2
Constant	5.4309	0.8855	6.133	0.0000	
BNMd_Tracks	1.0382	0.4173	2.488	0.0152	0.0343
BNMsq	-0.2078	0.1641	-1.266	0.2097	0.0089
Playlisted_Tracks	0.5422	0.2665	2.035	0.0457	0.0230
Average_Album_Score	0.2086	0.1182	1.765	0.0820	0.0173
Playsq	-0.1340	0.0800	-1.675	0.0983	0.0156
Total_Forkcast_Tracks	0.0117	0.0220	0.532	0.5966	0.0016
Year_End_Lists	0.0318	0.0399	0.798	0.4278	0.0035
Album_Number	-0.0216	0.0320	-0.675	0.5016	0.0025

Regression Equation: Album_Score = 5.43089552924 + 1.0382154035 BNMd_Tracks − 0.20780643438 E

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Statistic	P Value	ΔR^2
Regression	8	48.6771	6.0846	13.793	0.0000	0.6119
Residual	70	30.8794	0.4411			
Total	78	79.5565				

R-Squared: **0.611856**
Adjusted R-Squared: **0.567497**
Standard Deviation of the Error: **0.664179**

<new filter>

Figure 3: *Model of collection 1*

The most signficant predictor of an album's score according to the model is the number of tracks from that album that received the award of "Best New Music." When a linear regression t-test is perfomed, we observe a test statistic of 2.488 and a P-value of 0.0152. It makes

sense that this variable is most correlated with an album's score, since if Pitchfork rates tracks from a certain album highly, it is much more likely that they will favorably review the album itself. Also highly correlated is the average album score, with a test statistic of 1.765 and a P-value of 0.0820. This certainly makes sense, since artists that Pitchfork doesn't like are unlikely to change drastically between albums. In fact, I would have been somewhat surprised that this is not the most powerful predictor if not for the preponderance of recent large scale disappointments (The Strokes' Angles, MGMT's *Congratulations*, and Panda Bear's *Tomboy*) and surprises (Ariel Pink's Haunted Graffiti's *Before Today*). Another important factor is the number of tracks placed on Pitchfork's "Playlist," which has a test statistic of 2.035 and a P-value of 0.0457. This is important for the same reasons as the "Best New Music" tracks, but the higher honor of the Best New Music award is reflected in its higher coefficient. As expected the number of Forkcast tracks is less important since it is less discriminating, yet it does a good job of gauging the amount of buzz that smaller bands generate on music blogs. A variable that I expected to have a greater influence on the model was the most recent score an artist received. I figured that if an artist followed an upward or downward trend, it would be reflected in the artist's most recent album, but with a P-value of 0.9769, it seemed clear that this was not the case. Ultimately, I excluded this explanatory variable from the final model. I also expected the album number to be a more powerful negative predictor since artists that are producing their tenth albums usually run out of ideas and fail to appeal to any fans larger than their base of extremely devoted fans. Though the correlation was negative as predicted, the high P-value of 0.7695 suggests that this is most likely not statistically significant.

After observing that the model consistently overpredicted better-than-average albums and underpredicted average ones, I took an alternate approach and attempted to create two separate models based on the average album score of each artist's previous albums: artists with above average scores would be assessed in a separate model from those with

below average scores. However, the resulting models were inaccurate and unreliable. In the "above average" model, most predicted album scores hovered around around 8 to 8.5, while nearly all of the albums in the "below average" model remained in the 7 to 7.5 range regardless of any of the artist's other qualities, suggesting that placement into either of the groups determined the prediction far more powerfully than nearly any other variable. This caused the residuals of both models to be consistently erratic, leading me to reject these models in favor of the more accurate second model.

Another somewhat irrelevant but interesting finding was the distribution of Pitchfork ratings. Although I did not collect enough data to produce a completely reliable study on the distribution of Pitchfork scores, it is interesting to note that the sample average I obtained (only including the preliminary sample) is approximately 7.00. The histogram and boxplot in Figure 4 illustrate that the album scores are slightly skewed right as would be expected. A surprising number of albums approach the 8 range, yet they are not awarded the Best New Music rating as other albums are, some of which receive lower scores than non-BNM'd albums.

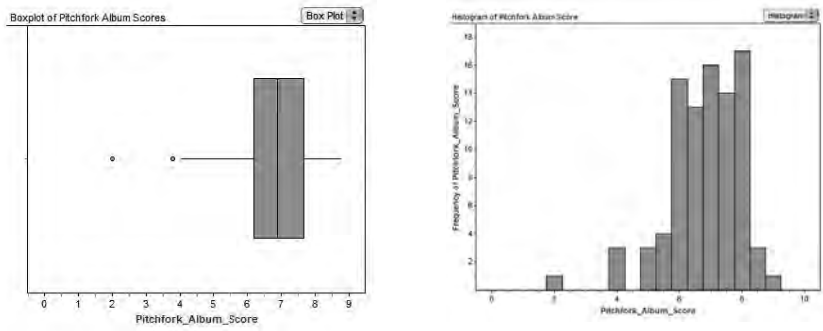


Figure 4: *Boxplot and historgram of album scores*

Applicable Albums Reviewed During the Week of April 25th	Album Score	Predicted Album Score (Final)	Predicted Album Score (Initial)	Residual (Final)	Residual (Initial)
Take Care, Take Care, Take Care (Explosions in the Sky)	7.2	6.4	6.8	-0.8	-0.4
Figurines (Figurines)	6.3	6.6	6.9	0.3	0.6
Liberty's Not for Eveyone (Chain and the Gang)	6.6	6.8	6.6	0.2	0.0
Music Sounds Better With You (Acid House Kings)	7.8	7.0	7.0	-0.8	-0.8
Terra (Julian Lynch)	7.7	7.8	7.8	0.1	0.1
Living on the Other Side (The Donkeys)	5.3	6.5	6.1	1.2	0.8
Swanlights EP (Antony and the Johnsons)	7.5	7.2	7.3	-0.3	-0.2
Circus (Black Devil Disco Club)	6	6.6	6.8	0.6	0.8
Canyon Candy (Javelin)	6.9	6.9	6.9	0.0	0.0
Cornershop and the Double-O Groove Of (Cornershop)	7.6	6.9	6.9	-0.7	-0.7
The Book of David (DJ Quik)	8.4	7.5	7.7	-0.9	-0.7
Losing Color (Stricken City)	7.7	7.0	6.9	-0.7	-0.8
Dancer Equired (Times New Viking)	6	7.7	8.2	1.7	2.2
The Only She Chapters (Prefuse 73)	6.4	6.7	7.0	0.3	0.6

Figure 5: *Albums reviewed week of April 25th*

In addition, I collected data from all of the albums released during the week of April 25th, none of which were factored into the either of the models. I input the data into each model and calculated the residu-

als. (See Figure 5.) In both models, most of the predictions were fairly close, but there were a few surprises. Times New Viking had tracks featured in the Playlist and a very consistent album history, but it was panned on release. On the other hand, Pitchfork showed few signs that they would award DJ Quik their Best New Music seal. For the most part, neither model appears to have a significant advantage over the other; although the final model produces slightly lower residuals in some cases, in nearly as many others, the residuals increased.

4 Conclusion

Despite its shortcomings, the regression I created to predict Pitchfork albums scores appears to be considerably reliable. Unsurprisingly, I discovered that the most powerful predictors of an album's score are the number of "Best New Music" tracks that are present on the album, the artist's average album score, and the number of tracks from the album that Pitchfork places on their Playlist. The non-linearity of the residual plot suggests that there were not enough influential variables that could significantly predict an album's critical failure or dampen the exaggerated effect of a couple of fantastic tracks. Many of the problems I encountered could be solved if I could spend more time to collect more data. My current model could be refined if I simply collected data spanning back further in Pitchfork review history. However, a more effective use of time would involve adding more variables, possibly some of which are unrelated to Pitchfork. For example, I could look at the effect of popularity by looking at the number of Last.fm plays an artist has during the weeks preceding their album release. I could also look at some of the popular indie music blogs or Hype Machine (a music blog aggregator) to measure blog buzz and more accurately evaluate underground popularity. It may also be worthwhile to look at an artist's genre of music to observe if Pitchfork has a tendency to favor any genre over another and see the effects of that bias on an album's score. It may also be effective to look at other professional music reviewers to see if there is a correlation between their reviews and Pitchfork's, even though they may not be the best predictors since few music sites release advance reviews and, perhaps more importantly, Pitchfork is usually the site that sets trends that other music reviewers and maga-

zines follow. Although it would be a painstaking process, accumulating and incorporating more explanatory variables would greatly increase the predictive power of the model I created.

Album Title (Artist)	Predicted Album Score
Sun & Shade (Woods)	7.7
Burst Apart (the Antlers)	8.2
Past Life Martyred Saints (EMA)*	8.2
Bon Iver, Bon Iver (Bon Iver)	8.5
Helplessness Blues (Fleet Foxes)	8.9
Cults (Cults)*	9.0
Gloss Drop (Battles)	7.4

*Using the preliminary model for an artist's first album release.

Figure 6: *Predicted album scores*

I'd like to end this paper with a response to my initial motivation to undertake the creation of this regression. What I really wanted to do was simply to find out what music is going to be worth listening to without having to wait for Pitchfork to release its review. So, I looked at a few noteworthy albums that are going to be released in the coming weeks and I plugged in their variables to predict their scores. Since some are weeks away from release, those variables may not be entirely accurate or up-to-date. Nevertheless, I'll have these albums on repeat for a while. And who knows? Maybe I'll be the one who will be able to say, "Cults? Fleet Foxes? EMA? I was listening to them way before they went mainstream." ●