Smells Like Statistical Spirit: Formulating a Rock and Roll Classic

Jonah Cox

2013

Smells Like Statistical Spirit: Formulating a Rock Album

This aim of this project is to examine the rock album *Nevermind* by Nirvana (1991), in order to analyze how much variance in the structure of the songs can be explained using regression analysis. The idea for this project was chosen to examine the truth behind the often opined view that rock music is formulaic, and the album was chosen because of its relatively simple melodic structures. Even band members of Nirvana have stated their songs followed a formulaic structure of "verse, chorus, verse", going so far as naming a song after the concept (Cobain, Noveselic, & Grohl, 2004), and it was of interest to see if this could be assessed mathematically. Of particular interest was to see if a model can be made which can inform what note comes next given relevant variables.

Method

To assess the greatest amount of variance that can be explained concerning the pattern of notes, it is important to decide how to detail the melodic structure of the album numerically. What other variables would be necessary to include, and how to adequately represent these details in spreadsheet form are also important factors in sussing out a statistical structure behind the album.

One data sheet was compiled with rather simplistic variables. These were: track number of song, first note of song, last note of song, previous songs first note, previous songs last note, length of song, and beats per minute of song (BPM). The data was compiled from all 12 songs on the album, excluding the first songs "previous songs first note", and "previous songs last note", as these data points do not exist.

Another data sheet was created with more in depth variables. These were: track number,

beats per minute (BPM) of song, length of song (in seconds), place in measure, current note in respect to place in measure, tonal distance to next note in sequence, tonal distance to previous note in sequence, next note in sequence, previous note in sequence, and dummy coding for verse, chorus, or bridge. This process resulted in 3687 lines of data.

In order to code the notes as a variable, a method similar to guitar tablature was utilized. In many formal music notations, "one step" consists of moving up or down relative to a scale; A to B, for example, which is a movement across two distinct notes; whereas moving a "half step" would be A to A sharp, which is a movement across one distinct note. Guitar tablature utilizes a reference to the fret number being played as a method of writing music, with 0 standing for the lowest note on a string. For the purposes of this assessment, a 0 was coded as the lowest note producible with E flat tuning on a guitar (the tuning method utilized by Nirvana). Each successive higher note possible was coded as a one point increase. The simplest melody was assessed throughout the album's songs. In some songs, the bass and guitar started playing the same melody, then one would diverge; the simplest melody was coded in these situations.

There were two songs on the album that employed "dropped tuning", or tuning one string of the guitar and bass down two notes, in order to create lower notes. I coded these lower notes where they occurred as negative numbers. This created an arbitrary zero; as the value of 0 was not truly the lowest note possible, but for ease of coding, the tradeoff was beneficial.

For the data to be accurate, the notation of the "BPM" and "place in measure" variable measurement had to be consistent across each of the individual tracks. In order to do this correctly, a metronome was synced with the notated beats of the measure. There can be some conjecture so far as when to start a new measure, but for consistency, an attempt was made to

hold the measure length at 8 beats across all songs. This was not possible in certain instances where the main melody timing of the verse was changed, with a denouement of half the length, or when the song resolved on one note at the beginning of a new measure. Also, in some songs, the simple melody flowed across multiple notes in spurts. In an attempt to keep with 8 notes in the melodic line, these were noted as decimal increments of whole beats. A dummy coded variable was used to denote if the note being assessed was part of the verse or the chorus.

Results

The first set of data to be analyzed was the simple set. Regression analysis was utilized to see what variables were beneficial in assessing the variable “first note”, or the first note of each song. When all the variables were ran in a model, the track number and last note variables were significant at the .05 level, and the previous songs first note variable was significant at the .10 level. All 6 variables combined as predictors resulted in an R squared value of .8298, with an adjusted R squared of .5746. This is a telling detail, that though a high level of variance was seemingly explained using the model, there was a significant amount of overlapping variance being explained by the variables utilized. The total F-statistic was 3.251 on 6 and 4 DF with a p-value of 0.1368.

Also, in compiling the data, it turns out that many of the songs resolve on the first note of the verses measure, which in turn is also likely to be the first note of the song. With this being the case, it is not surprising this variable would be statistically significant. When removed from the model, the significance drops dramatically. The overall model has no significant individual predictors, and the R squared reduces to .3188, with an adjusted R squared of -.3624. This model has an F-statistic of 0.4679 on 5 and 5 DF with a p-value of 0.7879.

When the variable “track number” alone was run as a predictor, it garnered an R squared of 0.3018, with an adjusted R-squared of 0.232. It is still significant, but reduced from the .05 level to the .10 level. Even so, this appears to be the most significant predictor of a songs first note, under the variables available in this data set. This one variable model has an F-statistic of 4.324 on 1 and 10 DF, with a p-value of 0.06427.

The larger data set created results of much higher significance. With all other variables ran as a predictor of the “next note” variable, the R squared value was .9811, with the adjusted R squared retaining the value of .9811. On reflecting why these results were so successful, it became apparent that coding a particular note, as well as “distance to next note”, creates an obvious indicator to what the next note would be; no regression model is necessary to obtain the next note when “distance to next note” is included as a variable.

A different model was created, removing the variable pertaining to the distance to the next note. The resulting model was still significant, with a p value < 2.2e-16. R-squared was 0.2304, with an adjusted R-squared of 0.2287. This is a large decrease in variance explained from the first model, but with the adjusted R-squared being close to the R-squared value, each variable is explaining variance with little overlap. Variables with large significance as predictors for the next note were track number, which was significant at a .05 level, with verse, note, and previous note significant to a near 0.0 level.

Discussion

There is an old joke about three statisticians out hunting. When a bird flies overhead, the first shoots ten feet too high, while the second shoots ten feet too low. The third exclaims “Direct hit!” This holds a lesson when analyzing data; it can be easy to lose track and miss the point of

exactly what statistical analysis can tell you about a data set, and what is beyond its scope. While the process of regression analysis can help establish the amount of variance that can be explained for one variable based upon other variables, the scope of this project obviously holds limited potential in the realm of establishing predictions based upon the variables, as it compiles information regarding an entire data set of one specific album, and is of no potential use predicting anything outside this album (i.e., there is no potential 13^{th} track from the album that can be predicted from the regression equation.)

While the simple data set initially looked promising, when it became apparent that the predictor variable pertaining to the "last note of song" was reliant on the variable "first note of song" to be predicted, and therefore removed from the model, the new model explained much less of the variance. When running the predictor variables individually, track number became the best predictor of "first note of song".

With the large data set, a successful model was created to predict the variable "next note" with a decent level of statistical significance, with nearly 23% of the variance in next note choice explained. Altogether, this last model did accomplish the goal set out in this proposal, which was to see how literally formulaic the album in question was. The conclusion that can be drawn is that it is somewhat formulaic.

Another variable that was not looked into specifically, but is perhaps worthy of further analysis pertains to era in which the album was created. The album was initially recorded in early 90's, at the tail end of vinyl record production, and during the heyday of cassette tapes. Cassettes and records are two sided, and many recording artists took this into account when crafting albums; in fact, the term "album side" refers to the flow of half of an entire recording on two

sided media. Looking at patterns in the albums construction as two separate 6 track segments might garner differing results, especially when using track number as a dependent variable.

References

Cobain, K., Noveselic, K., & Grohl, D., (1991). *Nevermind* [CD]. New York, New York: DGC.

Cobain, K., Noveselic, K., & Grohl, D., (2004). Verse Chorus Verse on *With The Lights Out* [CD]. New York, New York: DGC.

Appendix A

Output: Simple Data Set: 1st Model

```
Call:
lm(formula = first_note ~ track + last_note + prev_last + prev_first +
    length + bpm, data = firstlast.df)

Residuals:
     2      3      4      5      6      7      8      9     10     11     12
1.4462 -1.4298 -0.1397  0.6770 -2.7932  0.5288  1.2462  2.0711 -1.7527  1.2072 -1.0611

Coefficients:
             Estimate  Std. Error   t value Pr(>|t|)
(Intercept) -6.79930   11.05037  -0.615   0.5716
track        2.64113    0.84261   3.134   0.0350 *
last_note    1.27835    0.36881   3.466   0.0257 *
prev_last    0.51360    0.31233   1.644   0.1754
prev_first  -1.96106    0.80811  -2.427   0.0722 .
length       0.02816    0.03395   0.830   0.4534
bpm         -0.09366    0.05567  -1.682   0.1678
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.459 on 4 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8298,     Adjusted R-squared: 0.5746
F-statistic: 3.251 on 6 and 4 DF,  p-value: 0.1368
```

Output: Simple Data Set: 2nd Model

```
Call:
lm(formula = first_note ~ track + prev_last + prev_first + length + bpm, data = firstlast.df)

Residuals:
     2      3      4      5      6      7      8      9     10     11     12
5.7482 -3.5738 -0.6159 -2.7809 -2.3797 -0.1830  0.4118  1.1476  2.9327  3.3138 -4.0207

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.101152 19.683919 -0.158    0.881
track        0.188128  0.818476  0.230    0.827
prev_last   -0.081513  0.466938 -0.175    0.868
prev_first   0.382296  0.792276  0.483    0.650
```

```
length       -0.001101   0.058855  -0.019    0.986
bpm           0.037891   0.072890   0.520    0.625

Residual standard error: 4.401 on 5 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.3188,     Adjusted R-squared: -0.3624
F-statistic: 0.4679 on 5 and 5 DF,  p-value: 0.7879
```

Output: Simple Data Set: 3rd Model

```
Call:
lm(formula = first_note ~ track, data = firstlast.df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1154 -2.3007  0.1329  2.0507  4.5490

Coefficients:
             Estimate    Std. Error  t value Pr(>|t|)
(Intercept)   0.3182     2.0049      0.159   0.8771
track         0.5664     0.2724      2.079   0.0643 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.258 on 10 degrees of freedom
Multiple R-squared: 0.3018,     Adjusted R-squared: 0.232
F-statistic: 4.324 on 1 and 10 DF,  p-value: 0.06427
```

Output: Large Data Set: 1st Model

```
Call:
lm(formula = note ~ track_number + song_length + distance_next +
    distance_prev + measure_place + verse + note_next + note_prev,
    data = nirvana.df)

Residuals:
    Min       1Q   Median       3Q      Max
-6.32611 -0.06741  0.01388  0.08401  7.01764

Coefficients:
                   Estimate    Std. Error      t value Pr(>|t|)
(Intercept)       0.1602247  0.0701244    2.285  0.02238 *
track_number    0.0063908  0.0029821    2.143  0.03217 *
song_length    -0.0003648  0.0002268   -1.608  0.10784
distance_next  -0.7168568  0.0068714 -104.325  < 2e-16 ***
```

```
distance_prev  -0.2575308  0.0068426  -37.636  < 2e-16 ***
measure_place 0.0190562  0.0035267    5.403 6.95e-08 ***
versechorus     -0.0260218  0.0292065   -0.891  0.37301
verseverse       -0.0911024  0.0286666   -3.178  0.00150 **
note_next         0.7049501  0.0069972  100.748  < 2e-16 ***
note_prev         0.2648943  0.0070908   37.357  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4941 on 3675 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.9811,     Adjusted R-squared: 0.9811
F-statistic: 2.124e+04 on 9 and 3675 DF,  p-value: < 2.2e-16
```

Output: Large Data Set: 2nd Model

```
Call:
lm(formula = note_next ~ track_number + song_length + distance_prev +
    measure_place + verse + note + note_prev, data = nirvana.df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7598 -2.5816 -0.2285  2.2244 15.2401

Coefficients:
                   Estimate   Std. Error    t value Pr(>|t|)
(Intercept)        2.485718   0.445876   5.575 2.65e-08 ***
track_number    0.047283   0.019020   2.486 0.012966 *
song_length       0.002018   0.001449   1.393 0.163682
distance_prev   -0.069106   0.051345  -1.346 0.178413
measure_place  -0.016877   0.022584  -0.747 0.454942
versechorus        0.100973   0.186184   0.542 0.587625
verseverse         -0.902475   0.181942  -4.960 7.36e-07 ***
note                  0.268911   0.052765   5.096 3.64e-07 ***
note_prev          0.201335   0.052988   3.800 0.000147 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.157 on 3676 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.2304,     Adjusted R-squared: 0.2287
F-statistic: 137.6 on 8 and 3676 DF,  p-value: < 2.2e-16
```