MWSUG 2016 - Paper AA14

An Analysis of the Repetitiveness of Lyrics in Predicting a Song's Popularity

Drew Doyle, University of Central Florida, Orlando, FL

ABSTRACT

In the interest of understanding whether or not there is a correlation between the repetitiveness of a song's lyrics and its popularity, the top ten songs from the year-end Billboard Hot 100 Songs chart from 2006 to 2015 were collect. These songs then had their lyrics assessed to determine the count of the top ten words used. These words counts were then used to predict the number of weeks the song was on the chart. The prediction model was analyzed to determine the quality of the model and if word count is a significant predictor of a songs popularity. To investigate if song lyrics are becoming more simplistic over time there were several tests completed in order to see if the average word counts have been changing over the years. All analysis was completed in SAS® using various PROCs.

INTRODUCTION

With the goal of understanding whether or not there is a correlation between the repetitiveness of a song's lyrics and its popularity, the top ten songs from the year-end Billboard Hot 100 Songs chart from 2006 to 2015 were collected. These songs then had their lyrics assessed to determine the count of the top ten words used. These words counts were then used to predict the number of weeks the song was on the chart. The prediction models were analyzed to determine the quality of the model and if word count is a significant predictor of a songs popularity. Multiple models were tested to find the best possible model for predicting the number of weeks on the chart.

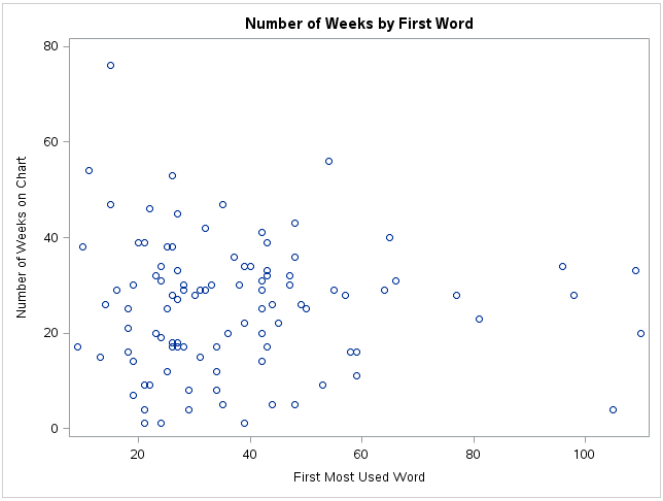
GETTING THE DATA INTO SAS

The first step is to correctly get your data into SAS. The first variable read in is Year for the year in which the song was on the year-end Hot 100 chart. The next variable is Weeks for the number of weeks the song was on the Hot 100 Chart. The next 10 variables represent the first most used word, the second, and so on until the ten most used word, named one to ten. Each word variables are assigned the total number of times that word is used in the song. It was done using the following code:

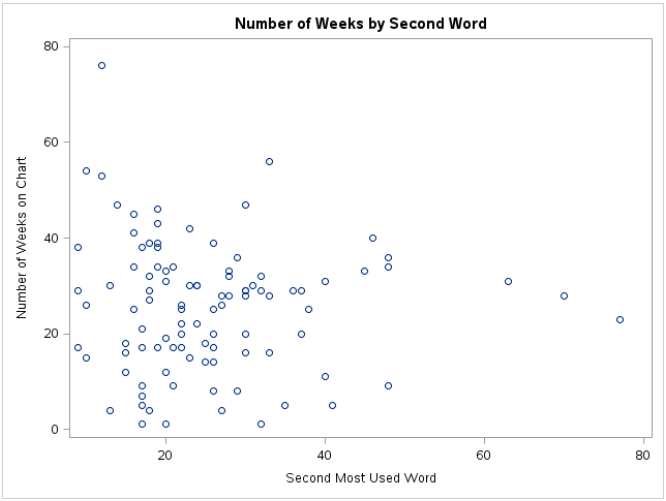
```
DATA Lyrics;
   INPUT Year Weeks One Two Three Four Five Six Seven Eight Nine Ten;
DATALINES;
2015 8 34 29 29 25 19 19 18 18 18 16
/*Rest of data*/
;
RUN;
```

SCATTER PLOTS

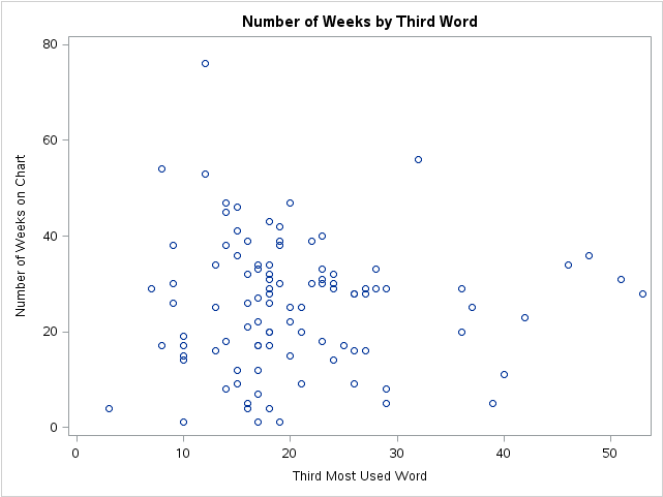
Scatter plots were created to understand the relationship between the dependent variable, Weeks, and the individual independent variables pertaining to the word counts.



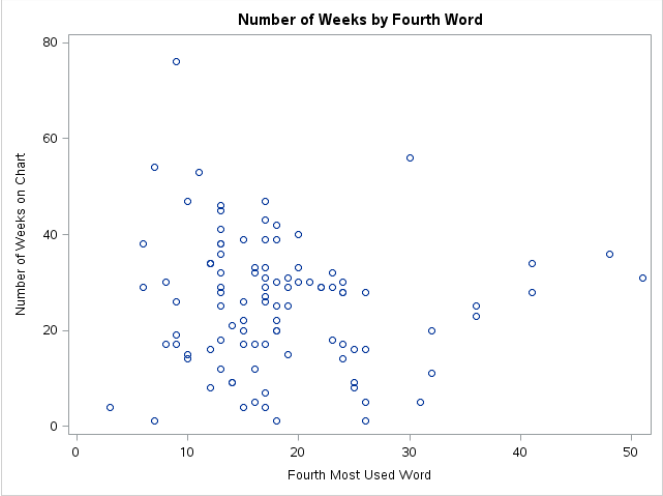
The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the first most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



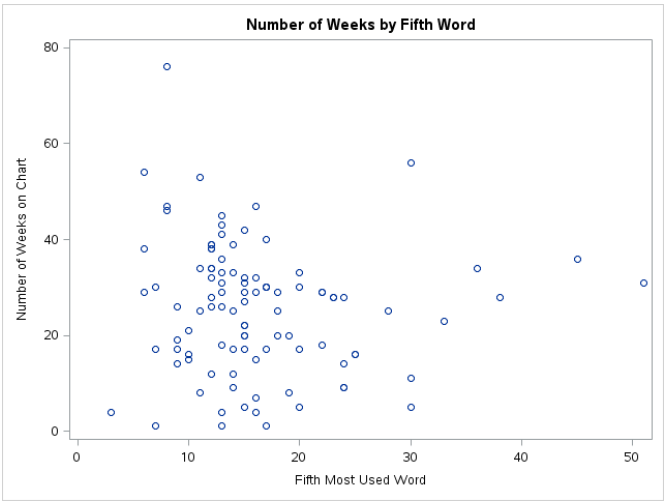
The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the second most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



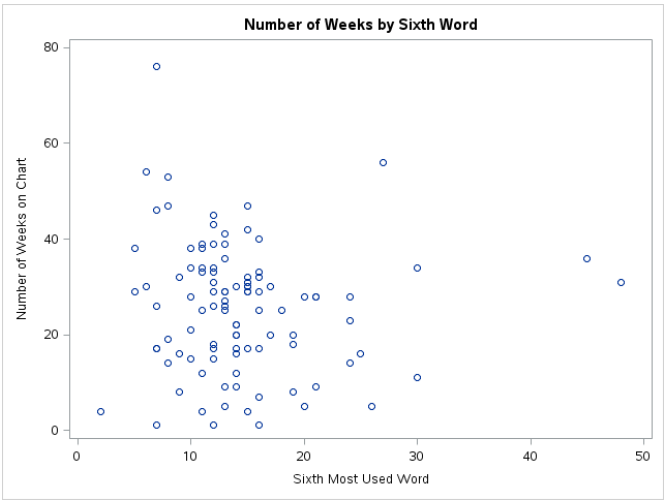
The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the third most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



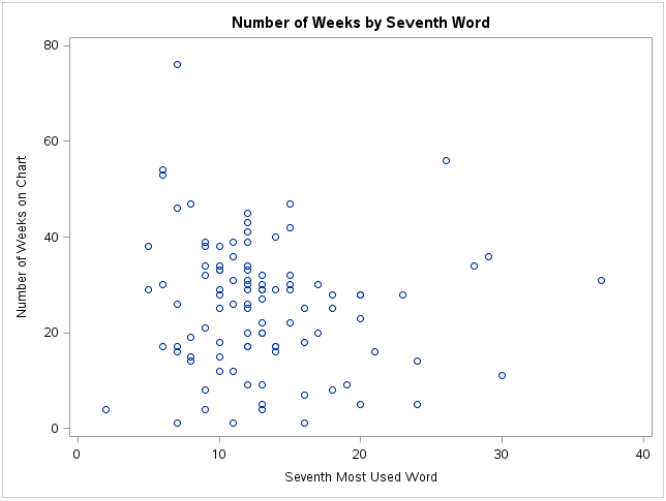
The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the fourth most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



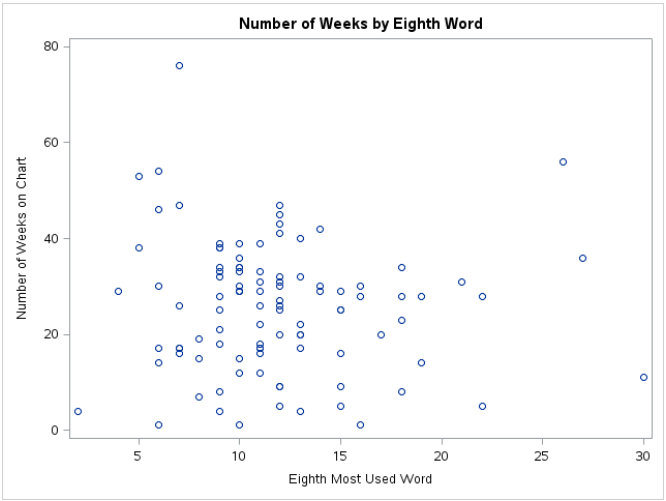
The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the fifth most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



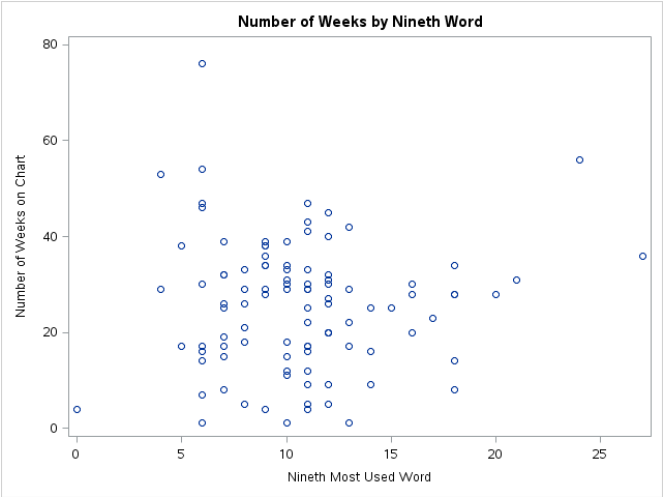
The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the sixth most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



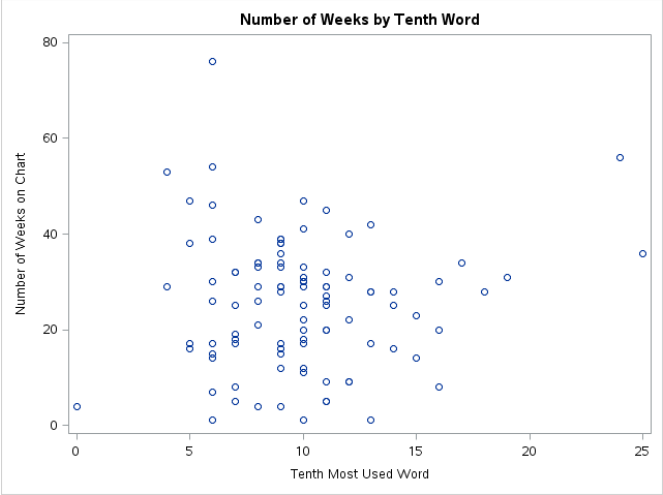
The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the seventh most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the eighth most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the nineth most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.



The plot above shows the relationship between the number of weeks on the Hot 100 chart and the count of the tenth most used word in each song. No clear trend is apparent in the plot, which suggests that there is no clear relationship between the two variables defined on the scatter plot.

From these scatter plots there are no clear trends between the number of weeks on the Hot 100 chart and the count of the most used words in each song. We will investigate this further to see if there is any possible correlation between these variables.

FINDING A GOOD MODEL

First, PROC REG was used on the base model containing the independent variables One-Ten. This was done to see the significance of the full model and each of the independent variables. No interaction between the variables were used at this stage. This is done using the following code:

```
TITLE "First Model: Everything included, no interaction";
PROC REG DATA= Lyrics;
```

```
    MODEL  Weeks = One Two Three Four Five Six Seven Eight Nine Ten;
RUN;
```

From the output below, one can see that the model is not significant at the significance level of 0.05 due to a p-value of 0.2986. It can also be noted that the independent variable Three is significant with a p-value of 0.0088 and all other variables are not significant. The R-squared for the model is only 0.1193 with an adjusted R-squared of 0.0203. This is not a good model for predicting the number of weeks on the chart.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	2176.28484	217.62848	1.21	0.2986
Error	89	16070	180.55624		
Corrected Total	99	18246			

Root MSE	13.43712	**R-Square**	0.1193
Dependent Mean	25.61000	**Adj R-Sq**	0.0203
Coeff Var	52.46826		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	24.76840	4.14686	5.97	<.0001
One	1	0.01138	0.10299	0.11	0.9123
Two	1	-0.24465	0.27128	-0.90	0.3696
Three	1	1.61847	0.60406	2.68	0.0088
Four	1	-1.25593	0.68384	-1.84	0.0696
Five	1	-0.88535	0.75829	-1.17	0.2461
Six	1	0.91601	0.90640	1.01	0.3150
Seven	1	-0.44129	1.07148	-0.41	0.6814
Eight	1	-1.11542	0.85805	-1.30	0.1970
Nine	1	0.62530	1.43833	0.43	0.6648
Ten	1	0.98632	1.49626	0.66	0.5115

Through the stepwise selection method, some possible models for this particular data will be chosen. Stepwise, backward, and forward selection will all be used to see what models they choose.

```
   PROC STEPWISE;
      MODEL Weeks = One Two Three Four Five Six Seven Eight Nine Ten /forward
backward stepwise;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	2128.93899	266.11737	1.50	0.1673
Error	91	16117	177.10825		
Corrected Total	99	18246			

Summary of Forward Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Seven	1	0.0070	0.0070	4.3448	0.69	0.4075
2	Ten	2	0.0363	0.0433	2.6755	3.68	0.0580
3	Three	3	0.0075	0.0508	3.9216	0.75	0.3872
4	Four	4	0.0285	0.0793	3.0403	2.94	0.0896
5	Eight	5	0.0106	0.0898	3.9741	1.09	0.2992
6	Two	6	0.0121	0.1019	4.7535	1.25	0.2663
7	Five	7	0.0056	0.1075	6.1884	0.58	0.4497
8	Six	8	0.0092	0.1167	7.2622	0.94	0.3338

The forward selection chose the model containing the variables Seven, Ten, Three, Four, Eight, Two, Five, and Six. The variable One and Nine were the only variable dropped from the complete model. From this table in the output, we can see the p-values for each one of the selected variables. Each has a p-value below an alpha of 0.50, this is because the forward selection uses an alpha of 0.50. Forward selection starts with no variables and adds variables one at a time. Most users do not use forward selection as their preferred method due to a low alpha level. This model was deemed not significant with a p-value of 0.1673.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1628.14139	407.03535	2.33	0.0618
Error	95	16618	174.92262		
Corrected Total	99	18246			

Summary of Backward Elimination						
Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
One	9	0.0001	0.1192	9.0122	0.01	0.9123
Seven	8	0.0016	0.1175	7.1757	0.17	0.6852
Nine	7	0.0026	0.1149	5.4432	0.27	0.6027
Six	6	0.0077	0.1072	4.2225	0.80	0.3729
Five	5	0.0058	0.1014	2.8051	0.60	0.4404
Two	4	0.0122	0.0892	2.0359	1.27	0.2619

The summary shown above is telling the user what variables were eliminated from the model. Therefore, the model that backward elimination chose contains Three, Four, Eight, and Ten. Backward elimination starts with the full model and eliminates one variable at a time until the best model remains. Backward elimination compares each variable's p-value to an alpha of 0.10, which is why this time more variables were eliminated from the model. This model was deemed significant with a p-value of 0.0618.

According to Stepwise selection, no variable met the 0.15 significance level to be included into the model. Thus, the model that stepwise selection chose contains no variables. The selection criteria were stricter in this model selection in comparison to the previous methods.

CHECKING CORRELATION

We will check to see if any of the variables are correlated with one another. This will help us to see if any of the variables should be included in the model as an interaction.

```
TITLE "Correlation between variables";
PROC CORR DATA = Lyrics;
    VAR    Weeks One Two Three Four Five Six Seven Eight Nine Ten;
RUN;
```

Each box gives the correlation coefficients between the two variables and below it the corresponding p-values. A small p-value tells us that the variables are correlated with one another. It turns out that all of the independent variables, One to Ten, are all correlated with one another. This was to be expected since certain words are typically used together and song titles are usually repeated often in a song. We will try interactions between the different independent variables to test if those models and variables are more significant.

INTERACTION MODELS

The chosen model from the backward elimination containing the variables Three, Four, Eight, and Ten will be used. This model was deemed significant at the significance level of 0.10. Interactions between these variables will be

tested to produce the best possible model. The following is a list of tested interaction models. This list is not the entire list of interaction models.

```
PROC GLM DATA = Lyrics;
    MODEL  Weeks = Three Four Eight Ten Three*Four*Eight*Ten;
RUN;
```

This model is significant with a p-value of 0.0196 at the 0.05 significance level. All of the variables are also significant at the 0.10 significance level. The R-Squared value is 0.131343 and a root MSE of 12.98499.

```
PROC GLM DATA = Lyrics;
    MODEL  Weeks = Three Four Eight Ten Three*Four*Eight;
RUN;
```

This model is significant with a p-value of 0.0352 at the 0.05 significance level. All of the variables are also significant at the 0.10 significance level. The R-Squared value is 0.117859 and a root MSE of 13.08538.

```
PROC GLM DATA = Lyrics;
    MODEL  Weeks = Three Four Eight Ten Three*Four*Ten;
RUN;
```

This model is significant with a p-value of 0.0237 at the 0.05 significance level. Only Three, Four, and Three*Four*Ten are significant at the 0.10 significance level. The R-Squared value is 0.126967 and a root MSE of 13.01766.

```
PROC GLM DATA = Lyrics;
    MODEL  Weeks = Three Four Eight Ten Four*Eight*Ten;
RUN;
```

This model is significant with a p-value of 0.0154 at the 0.05 significance level. All of the variables except for Ten are significant at the 0.10 significance level. The R-Squared value is 0.136654 and a root MSE of 12.945221.

```
PROC GLM DATA = Lyrics;
    MODEL  Weeks = Three Four Eight Ten Three*Four;
RUN;
```

This model is significant with a p-value of 0.0522 at the 0.10 significance level. All of the variables except for Three, Eight, and Three*Four are significant at the 0.10 significance level. The R-Squared value is 0.108417 and a root MSE of 13.15523.

```
PROC GLM DATA = Lyrics;
    MODEL  Weeks = Three Four Eight Ten Three*Eight;
RUN;
```

This model is significant with a p-value of 0.0537 at the 0.10 significance level. All of the variables except for Three and Three*Eight are significant at the 0.10 significance level. The R-Squared value is 0.1077753 and a root MSE of 13.16012.

```
PROC GLM DATA = Lyrics;
    MODEL  Weeks = Three Four Eight Ten Three*Ten;
RUN;
```

This model is significant with a p-value of 0.0325 at the 0.05 significance level. All of the variables except for Three, Eight, and Ten are significant at the 0.10 significance level. The R-Squared value is 0.119734 and a root MSE of 13.07147.

```
PROC GLM DATA = Lyrics;
    MODEL  Weeks = Three Four Eight Ten Four*Eight;
RUN;
```

This model is significant with a p-value of 0.0339 at the 0.05 significance level. All of the variables are significant at the 0.10 significance level. The R-Squared value is 0.118710 and a root MSE of 13.07907.

```
PROC GLM DATA = Lyrics;
      MODEL  Weeks = Three Four Eight Ten Four*Ten;
RUN;
```

This model is significant with a p-value of 0.0233 at the 0.05 significance level. All of the variables except for Eight and Ten are significant at the 0.10 significance level. The R-Squared value is 0.127374 and a root MSE of 13.01462.

```
PROC GLM DATA = Lyrics;
      MODEL  Weeks = Three Four Eight Ten Eight*Ten;
RUN;
```

This model is significant with a p-value of 0.0177 at the 0.05 significance level. All of the variables except for Ten are significant at the 0.10 significance level. The R-Squared value is 0.133552 and a root MSE of 12.96847.

After considering all of the models, the model containing Three, Four, Eight, Ten, and the interaction between Four, Eight, and Ten was chosen as the best model. This model had the smallest p-value of all of the models; only one variable was not significant, largest R-square value, and smallest Root MSE.

NORMALITY

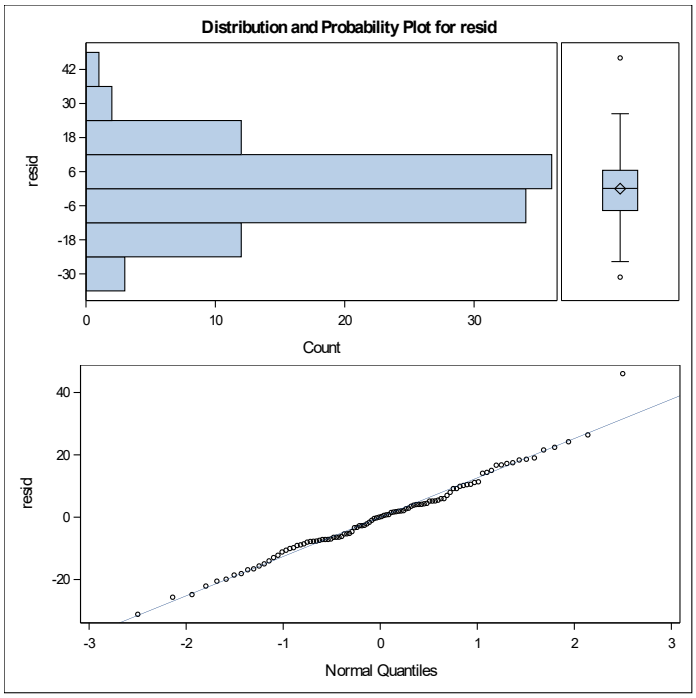
We want to test to see if the residuals of our selected model are normally distributed. Using the code below, we can look at the hypothesis test for normality and the distribution/plot of the residuals.

```
PROC GLM data = Lyrics;
    model  Weeks = Three Four Eight Ten Four*Eight*Ten;
    output out = new
    residual = resid ;
RUN;

PROC UNIVARIATE normal plot;
    VAR resid ;
RUN;
```

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.985113	**Pr < W**	0.3236
Kolmogorov-Smirnov	**D**	0.068499	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.055062	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.324067	**Pr > A-Sq**	>0.2500

According to both the Shapiro-Wilk and Kolmogorov-Smirnov tests for normality, we can say that the distribution of the residuals is normal. Both produce a test statistic with a corresponding p-value great that an alpha of .15, which means we cannot reject the null hypothesis that the residuals are normally distributed.



Next, we look at the distribution and probability plot for the residuals to also check for normality. We can see that both the histogram and boxplot are normally distributed. The points on the probability plot should form a linear shape. On the chart above, the points do form a linear shape excluding the singular outlier point.

CONCLUSION

Overall, the assumptions for the analysis of our chosen model held. Several different models were considered to see if the word counts of a song's lyrics could, to a certain degree, predict the number of weeks the song would remain on the Hot 100 Billboards Chart. With this analysis, there is not enough sufficient evidence to prove the hypothesis that song lyrics can predict the number of weeks on the chart. This analysis also indicated that the word count was not the best predictor of weeks on the chart. Due to the low R-Squared value of the best model and overwhelming amount of sub-par models with even lower R-Squared values, we can say that

Even though a model was discovered that met the requirements, there is still not enough evidence to disprove the initial hypothesis that the word count of a song's lyrics can predict the number of weeks the song would remain on the chart. There is evidence of correlation, but it was not irrefutable evidence to ascertain the confidence needed to prove or disprove the hypothesis. There is an opportunity for further study into this topic.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Drew Doyle
Enterprise: University of Central Florida
E-mail: drewdoyle@knights.ucf.edu