Predicting Speaker Through Speech Patterns: An application to the *Dear Hank & John* podcast

APMTH 101 Final Project

Egemen Bostan
Alexander Mancevski

Submitted on:
11 December 2020

Modified in December 2021 to remove personal details and to improve presentation.

Preamble

Dear Hank & John is a comedy podcast hosted by Hank and John Green about death, AFC Wimbledon (a third tier football team from England), and Mars ("Dear Hank & John"). Apart from providing dubious advice to its listeners, this podcast nurtures a community that is willing to create interesting projects about the podcast. One such experiment was the investigation of the ratio of its 2 hosts' talking duration (Dressel and Dressel). This project uses and builds upon the data created by Maggie and Peter Dressel to execute multiple regression analysis. You can find our raw data and analysis methods in Appendix A.

Model

In this model, we are looking to answer the following question: *Can we predict the ratio of Hank's talking time to John's talking time, given that we know different (listed) variables about the podcast episode?*

Our initial model has the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

where:

y = Hank's talking time / John's talking time : "Hank ratio"

x_1 = Episode Number

x_2 = Mean Frequency of the episode: "Mean frequency"

x_3 =Average Loudness of the episode: "Loudness"

x_4 = Frequency of the word 'Mars'

Variables and Data Collection

The aggregate raw data and relevant sources and methods are in Appendix A. Below is a description and discussion of the variables.

Independent Variable y : Hank talk-minutes per 1 John(other host) talk minute

➔ Data Source : https://docs.google.com/spreadsheets/d/1GCmbkYCQn-2sCkSUXmBE7X3sM2b9X3avmxhWGBvcDZw/edit?usp=drive_web&ouid=104251441230285677880. Please see Appendix A for a more complete explanation of the data collection process.

Regressor x_1: Episode Number (ordered by the date of release, first is earliest)

➔ *Data Source*: https://nerdfighteria.info/c/dearhankandjohn/transcribed

➔ *Justification*: As the podcast goes on, Hank - the less sociable and experienced of the podcast hosts - gets more comfortable with the medium and talks more. So, we expect a positive (potentially logarithmic) relationship.

Regressor x_2: Average Frequency of the episode recording

➔ *Data Source:* https://drive.google.com/file/d/1xx_E1wzOFSgy76etHuN7oYntYQI--Kcy/view?usp=sharing. The average frequency was calculated from the .mp3 file for the podcast episode, and represents the average pitch of the recording.

➔ *Justification:* Hank's voice is generally higher in pitch. He also expresses strong emotions more frequently, which produce higher pitched sounds. The more he talks (relative to John), the higher the average frequency will be. So, we expect a positive linear relationship.

Regressor x_3: Average loudness of the episode

➔ *Data Source*: https://drive.google.com/file/d/1xx_E1wzOFSgy76etHuN7oYntYQI--Kcy/view?usp=sharing. Calculated from the .mp3 file for the podcast episode by averaging the signal amplitude of the recording.

➔ *Justification:* Hank's talking volume is higher than John's. The more he talks, the greater the loudness of the episode will be. So, we expect a positive linear relationship.

Regressor x_4: Frequency of the word Mars in the podcast occurrences of the word "Mars"

➔ *Data Source:* https://nerdfighteria.info/c/dearhankandjohn/transcribed . Calculated by finding the proportion of the word 'Mars' in the episode transcript.

➔ *Justification:* Hank is interested in space, especially Mars (he presents a bit on Mars every episode). Conversations involving the word Mars are where Hank can contribute more than John (other host). Thus, we expect a positive linear relationship.

Data Analysis

We plot each regressor variable against the independent variable in Figure 1..

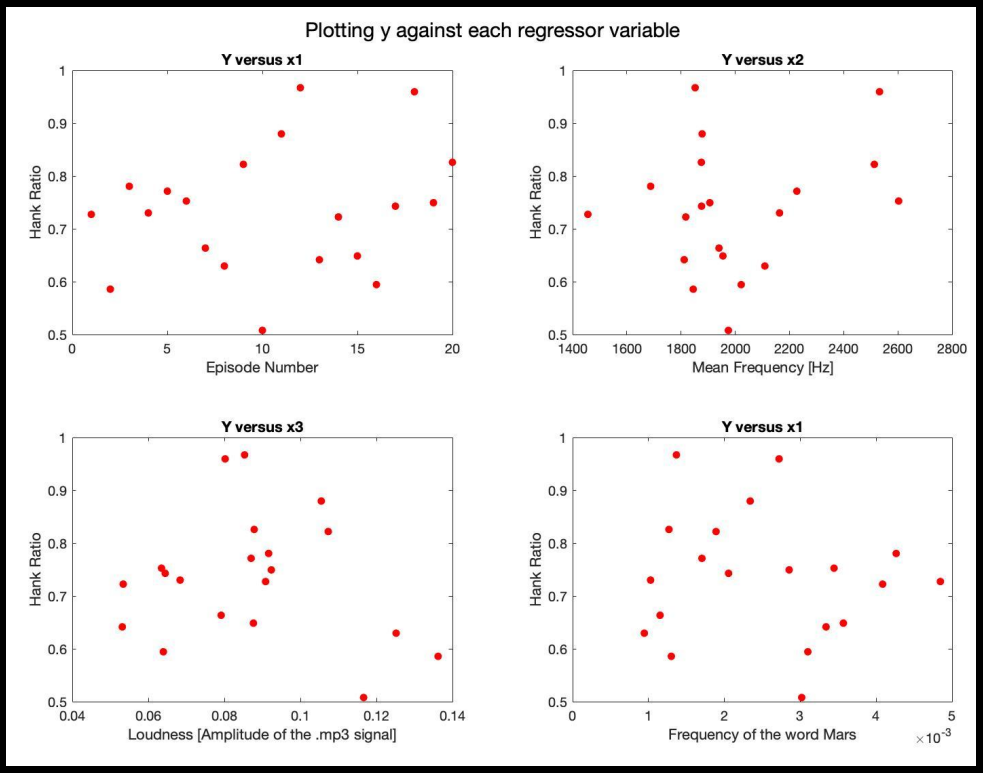


Figure 1: Regressor vs Independent variable Plots

None of the plots in Figure 1 indicate a strong linear relationship between the regressor variable(s) and the independent variable. Thus, we anticipate failing to reject the null hypothesis in the regression analysis.

While we haven't found a strong link between the variables, we can still investigate the regressor variables within themselves to check if there is an underlying variable driving both (or several) regressor variables. To check for collinearity, we plot regressor variables against each other.

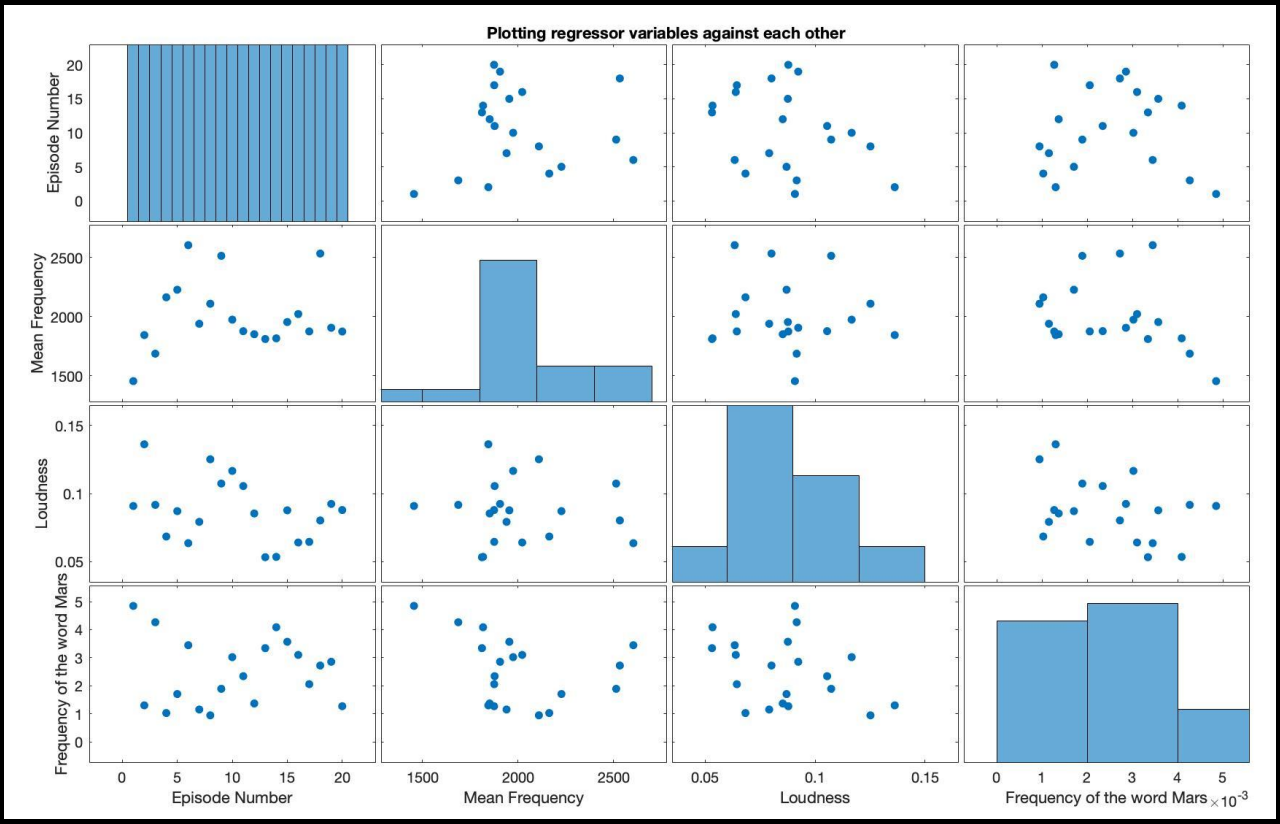


Figure 2: Checking for Collinearity by plotting regressor variables against each other does not find collinearity.

None of the plots in Figure 2 follow roughly straight lines. Thus, visual inspection suggests that there is no linear relationship between any two regressor variables. In other words, we do not observe collinearity.

Normally, in the case of established linearity and/or collinearity, outliers would be discussed. However, since there is no evident linear relationship between the independent variable and the regressor variables, it is very challenging to determine potential outliers. As the data collection method was consistent and the plots in Figures 1 and 2 do not show any data point exceedingly distant to the rest, we conclude that there are no outliers.

MLR Parameter Estimation and Confidence Intervals

- β_0: 0.062001
- β_1: 0.55109 episodes
- β_2: 0.49796 Hz
- β_3: 0.63549 *the amplitude of the .mp3 signal has no unit
- β_4: 0.72473 *no unit since [#word/#allwords]

The standard deviation of the residuals is 0.1270 and the standard deviation of the y-values

is 0.1195, which tells us that there is not a strong linear relationship. standard deviation of the residuals is larger than the standard deviation of the y-values. If there were to be a stronger linear relationship, then we would expect the standard deviation of the y-values to be larger than the standard deviation of the residuals.

Test for Significance of Regression

Using the F-test, we see that the linear relationship is not valid, with F_0 =0.9897, and $f_{critical}$ =5.9781. We fail to reject the null hypothesis.

Final Model Building

Based on the results above, none of the regressor variables are fit to model this relationship since all p values are greater than type 1 error, α, of 0.05. Furthermore, we considered increasing alpha to accommodate a weaker correlation inside the model, but all p values are much greater than alpha. None of the variables show strong or moderate evidence for linear relationship with the independent variable. This conclusion agrees with the initial observation where the plots of independent variables against regressors showed no linear distribution of points.

```
Number of observations: 20, Error degrees of freedom: 15
Root Mean Squared Error: 0.127
R-squared: 0.109,  Adjusted R-Squared: -0.129
F-statistic vs. constant model: 0.457, p-value = 0.766
Initial columns included: none
Final columns included:  none
    {'Coeff'     }    {'Std.Err.'  }    {'Status'}    {'P'     }
    {[    0.0044]}    {[    0.0046]}    {'Out'   }    {[0.3518]}
    {[9.4823e-05]}    {[9.5403e-05]}    {'Out'   }    {[0.3334]}
    {[   -0.8484]}    {[    1.2169]}    {'Out'   }    {[0.4946]}
    {[  -11.1614]}    {[   23.6894]}    {'Out'   }    {[0.6432]}
```

Figure 3: Stepwisefit function output for the dataset

The conclusion that none of the variables correlate is further confirmed by running the `stepwisefit` function in matlab, which comes to the same conclusion.

However, the sections of this assignment are pre-determined. So, for the purpose of creating a final model that yields itself to the analysis we're expected to fulfill in this section, we've decided on a new model.

Among the tested regressors, we see that x_2, mean frequency, has the lowest p-value. We also believe that this variable has the strongest justification for creating a linear relationship with the independent variable. While both hosts can talk louder, can talk about Mars, and can talk more in later episodes, the pitch of one's voice is difficult to change. So, we recognize that the mean frequency variable has the greatest potential to correlate with Hank's ratio (y), even if our dataset shows the evidence for this relationship is not sufficient.

Running regression analysis for the model $y = \beta_0 + \beta_1 x_2 + \varepsilon$, we get the following parameter estimates and p-values.

Table 1: Final Model Parameters

	Estimate	SE	tStat	pValue
(Intercept)	0.54519	0.19289	2.8265	0.011183
Mean Frequency	9.4823e-05	9.5403e-05	0.99392	0.33343

which yields $\beta_0 = 0.54519$, $\beta_1 = 9.4823 \times 10^{-5}\ seconds\ [1/Hz]$. P-values corresponding to the beta values are presented in the last column of the table.

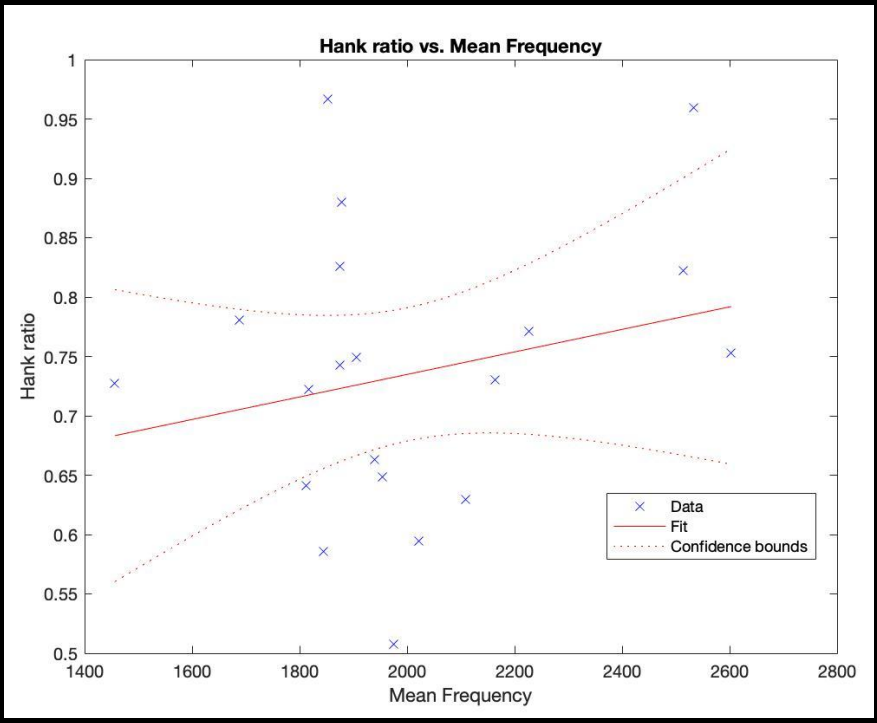


Figure 4: Independent variable versus mean frequency plot in detail shows the limitations of the linear fit for this variable.

The F-test is run on the final model and R^2 and adjusted R^2 values are extracted from the model. The condition to reject the null hypothesis is $F_0 > f_{critical} = f_{0.975,1,n-2} = 5.9781$. The calculated value is $F_0 = 0.9879$. Hence we conclude that regression is not significant, as expected. Similarly, the R^2 and adjusted R^2 values are $R^2 = 0.0520$ and adjusted $R^2 = -6.3817 \times 10^{-4}$. The values for R-squared parameters do not agree even if this is a 1 regressor model because Matlab treats the intercept coefficient β_0 as a regressor coefficient,too. The R^2 value also comes to the same conclusion- showing that the data presents very weak evidence for correlation.

At a confidence level of 95%, the confidence intervals for the parameters are as follows.

Table 2: Confidence Intervals of Final Model Parameters

Parameter	Lower Bound	Upper Bound
β_0	0.1400	0.9504
β_1	-0.0031	0.0003

The confidence interval for β_1 includes 0, which also would lead us to conclude that we cannot determine that there is a linear relationship. As before, all parameters agree on this result.

Analysis of Residuals

Our analysis of residuals will focus solely on x_2, as it is the subset of variables that best describes the data set.

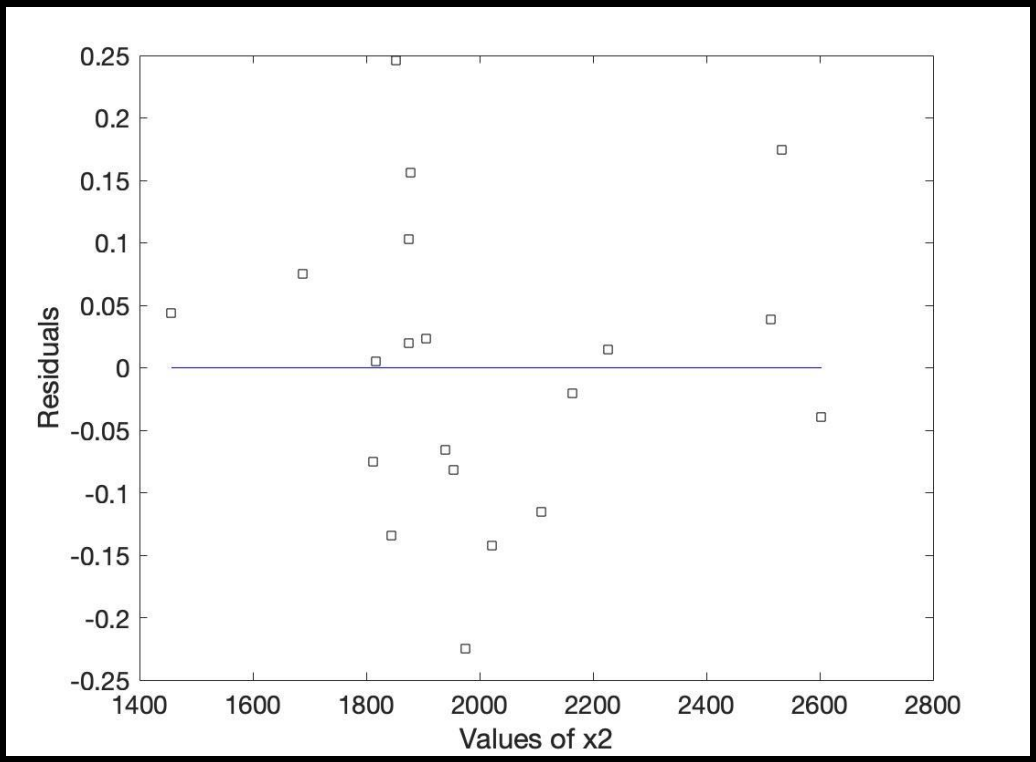
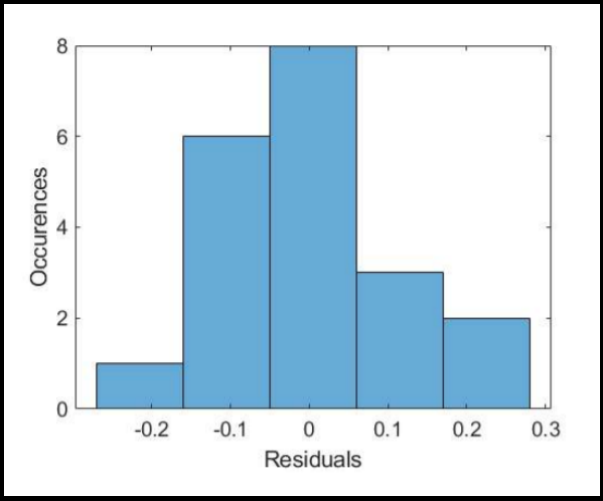


Figure 5: Residuals plot for x_2



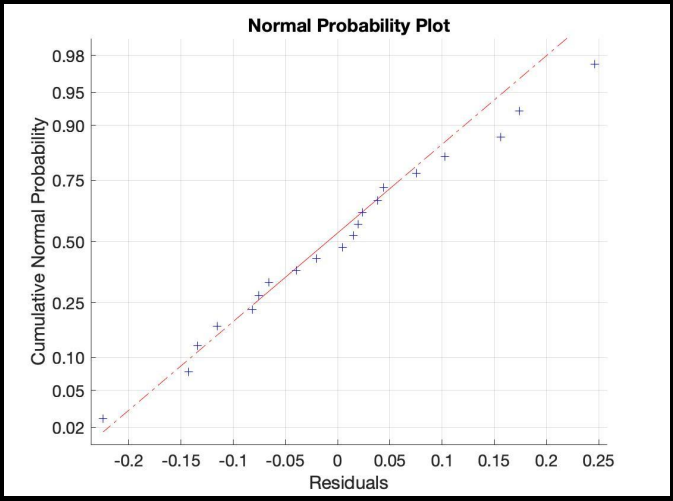


Figure 6: Histogram and Normal Plots for Residuals

The residuals follow the assumptions of a normally distributed, zero mean, constant variance random variable.

Discussion

We set out to answer the question: *Can we predict the ratio of Hank's talking time to John's talking time, given that we know different variables about the podcast episode?* The answer, according to the variables he had picked and data that we analyzed, is no. None of the listed variables strongly or moderately correlated with Hank's talking time ratio.

Having concluded this, we picked the mean frequency variable as the only regressor variable in the final model, as it had the strongest justification and had the lowest p value. This model represents "the strongest these variables can correlate", even if there is not still very strong evidence for linear regression.

With the final model, we obtain $R^2 = 0.0520$and $p = 0.333$. The low R^2 value indicates that the evidence for a linear relationship is weak and the low p value (significantly above 0.05) shows that there is not enough evidence to conclude a non-zero correlation between the regressor and the response variable exists. This is apparent from the distribution of the datapoint in the 4(a) figure.

The multiple linear regression model has a set of assumptions that must be applicable. Here, we discuss their applicability to the chosen model.

Linearity: The predicted relationship between reponse (y) and each regressor variable is linear. For any of the variables, there is no reason why the effect of an increase in regressor depends on the current value of that regressor, so the linearity assumption is satisfied for the model.

Normal distribution of residuals: The analysis in section 5 confirms the validity of this assumption for the final model and similarly for all the dataset in section 2.

Independence of observations: The model assumes that within a regressor variable, each data point is independent of the other. For episode number, it is clear that each episode number increases by 1, so the assumption is valid. For mean frequency, the frequency of one datapoint cannot affect another's, so the assumption is valid. For loudness, the relative loudness of a previous episode might have caused the editors to tweak the loudness of the next one, so this assumption may not hold. For the frequency of the word Mars, Hank talking more about Mars might have caused John to talk more about Mars, so some observations may have affected others, so this assumption may not hold.

The variance of residuals is the same for any value of regressors: There is an argument to be made that while the Hank ratio works for this model between 0.4 and 1.5 or 2, it will not linearly increase at its bounds. If Hank talks 90 percent of the time, the ratio would be 9. This is highly unlikely, yet if that were to happen, the variance of residuals would potentially increase, since any factor affecting Hank's talking time proportionally causes a greater chance numerically, hence the residual variance widens at high values. We conclude that in our dataset, the assumption holds, but we would not extrapolate beyond a ratio of 1.5 or 2.

No collinearity: This assumption was satisfied though the investigation in section 1.

Considering that our investigation concluded that none of the tested factors impacted the response variable, we cannot discuss predictive abilities of our model. However, this result helps us to negate the "stereotyping" that led to the justifications. While Hank is a self-admitted "space nerd", he does not overwhelm the conversations about Mars - John still contributes, even though he refers to Mars as a 'cold, dead rock far away in space'.

Similarly, we see that while Hank sounded to me to always talk louder, it is not the case. This suggests that it is not high volume but intonation or other characteristics of his voice that causes this effect. This recommends further investigation into understanding what causes us to hear something as louder, even when it isn't.

Furthermore, we see that the mean frequency does not correlate with Hank's talking ratio. Upon investigating the data source, we realized that in 6 episodes, a guest is hosted in the podcast and that 5 out of 6 times, the guest was female - so more likely to have a higher pitched voice than either John or Hank. We believe that if we had access to a bigger dataset, we would find a statistically significant correlation between this variable and the response variable. We invite others to obtain more data to evaluate this speculation.

Also, the episode number variable did not correlate with Hank's talking time ratio. While the initial justification seemed reasonable, we ignored the fact that both John and Hank have been science communicators and hosts of large conferences, so they must have been both comfortable with talking in this format. As a result, we were able to recognize that our initial hypothesis for this relationship was not valid. However, given that the podcast currently has 200+ episodes, it would be interesting to see if Hank ratio converges to 1, or how it evolves over time (“Dear Hank & John”). We do not think it is right to take all 200+ episodes and analyze for Hank ratio response variable, because our model cannot be extrapolated to large values of Hank ratio, as discussed in 6(c).

To sum up, while our model failed to find strong evidence for linear relationship between our regressor variables and response variable, we were able to learn about the dynamics of the conversation that takes place in Dear Hank and John.

Workload Distribution

Egemen: Data Collection, MATLAB Script, Discussion
Alexander: Preamble, Data Analysis, Residual Analysis

Sources

"Dear Hank & John." *Complexly*, https://complexly.com/shows/dear-hank-john. Accessed 9 Dec. 2020.

Dressel, Maggie, and Peter Dressel. *Dear Hank and John Time Distribution*. University of Iowa, https://docs.google.com/document/d/1z7wkCfhEQDA-3PDuo5o6nrC70tVWsH_-wX-5YbgfxVw/edit?fbclid=IwAR3GtbT7Q-IiD5YqL0LWd1hqWMCk_Q4G37d2S5hE2a665Dq8QmRmVgeVmaE. Accessed 9 Dec. 2020.

Appendix A - Raw Data Table

y	x1	x2	x3	x4
0.7273153153	1	1456.409849	0.09084109	0.004845814978
0.5857292451	2	1844.987834	0.13622412	0.001301066875
0.7804361487	3	1687.978581	0.09165101	0.004263451926
0.7300349606	4	2163.438414	0.06835742	0.00102739726
0.7711962834	5	2227.143624	0.087039374	0.00170440711
0.7525773196	6	2602.949359	0.063467346	0.003444733184
0.6634980989	7	1940.221855	0.07915076	0.001152516327
0.6295460953	8	2109.34375	0.12519069	0.0009449360021
0.821969697	9	2513.21209	0.1073303	0.001890148988
0.5077915099	10	1974.694102	0.11663898	0.003019165135
0.8796212121	11	1877.932805	0.10549688	0.002340207693
0.9669908336	12	1852.072309	0.08532049	0.001367573317
0.6414874142	13	1811.755009	0.0531487	0.003339567192
0.722410626	14	1817.488141	0.053390946	0.004085237938
0.6485809683	15	1954.723993	0.087641194	0.003567508233
0.5943804483	16	2022.094744	0.063955754	0.00309961522
0.7428357938	17	1875.772291	0.06444284	0.00205497046
0.9594459355	18	2532.407206	0.08020495	0.002720559658
0.7493540052	19	1906.321364	0.09235335	0.002853519341
0.8258669274	20	1875.004537	0.087846555	0.001269621422

Note: More information about the methods used in collecting the data can be found in the following spreadsheet and the python notebook.

Data collection spreadsheet:
-contact via email-

Python Notebook:
-contact via email-