PREDICTING THE POPULARITY OF TED TALKS USING COMPOSITE MEASURES OF POPULARITY AND MIXED MODELS

Marcos Cardenas Zelaya

March 15, 2020

ABSTRACT

Objectives: Determine what predicts the popularity of TED talk videos using average views per day and a composite measure of popularity.

Design: Poisson and Linear Mixed-Model Regressions are used to determine the popularity of TED talk videos and whether the popularity varies by when the video was published and its theme.

Data: Web scraped data from the TED talks website. The data contains descriptions of videos created during 2006 to September 21st, 2017. The original data set contained 2550 observations.

Predictors and Response: Average views per day and a composite measure of popularity that encompass number of languages the video has been translated to, number of comments, a string dictionary of the ratings given to the talk (e.g., inspiring, fascinating, jaw dropping, etc.) and their frequency, and the number of related talks. The predictors in this analysis are a talks duration, number of speakers in the video, how old the video is, when the video was published, the ”label” of the title and themes associated with the video, and the length of the title.

Results: We find that popularity varies by when it was published and its by its theme. Further, talks duration, talks film age in days since published, and the talks age group (when the talk was published) are the strongest predictors of popularity.

Conclusion: Although there is variation among themes, a talks duration, talks film age in days since published, and the talks age group (when the talk was published) are the strongest predictors of popularity. This suggests that the best way to create a popular video is by frequently publishing long videos, regardless of the content. Popularity seems to be a function of quantity not quality.

1 Introduction

TED is a non-profit organization that spreads ideas, primarily via short talks that can be accessed on the internet such as YouTube. The talks are conference like in style, ranging from themes on science to business to global issues. TED talks has an incentive to produce high quality videos that garnish views to maintain current viewers and attract new viewers. Finding what characteristics predict a talks popularity/success is essential to TED talks growth.

2 Data

We used descriptive data on TED Talk videos that has been web scraped from the TED talks website. The data contains descriptions of videos created during 2006 to September 21st, 2017. The original data set contained 2550 observations. Each observation includes descriptions of when and where the video was filmed, when it was published, who is/are in the talk, how many comments and views the video has obtained, the title of the video, duration of the video and other variables to describe the video.

3 Response and Predictors

3.1 Responses

In this analysis, we used two measures of popularity; the most intuitive being a form of views count, and a composite response. Table 2 summarizes the two responses.

3.1.1 Average Views Per Day

The most intuitive measurement of popularity is the number of views. However, the flaw of this is that it does not account for age of a video, that is an older video will have more views than a newer video just by the nature of being around longer and able to garnish more views. Indeed, this raw measurement would suggest that a day old video with 1 million views is just as popular as a 5 year old video with 1 million views. To account for this, we divide the number of views of a video by the number of days since it has been published. This allows videos to be comparable across length since publication. Number of views are in thousands.

3.1.2 Popularity

Besides number of views, the data set included the number of languages the video has been translated to, number of comments, a string dictionary of the ratings given to the talk (e.g., inspiring, fascinating, jaw dropping, etc.) and their frequency, and the number of related talks. We composed a composite popularity score using these variables. We use equal weightings in the construction of the composite variable, however unequal weight could be given if a prior knowledge of a particular variable should be weighted more [1]. The intuition for including each is provided as:

- **Number of Languages**: A ”popular” talk will be translated into several languages as their is a great demand for the talk.
- **Number of Comments**: A ”popular” talk will garnish an active comment section as people discuss/praise the video. We assume that a popular video will have many comments whereas an unpopular video will have few comments as individuals are less likely to finish watching the video and thus not comment. Further, we distinguish videos that generate lively comments and controversial comments. Lively comments include praise for the videos, whereas controversial comments will result from viewers debating the topic as the topic could be unpopular but controversial (i.e religion, politics etc...). To account for this we use a comments per views metric instead to make comments proportional to the number of views so that we can capture popular videos with a lot of comments and not unpopular/controversial videos with few views but a lot of comments.
- **Ratings**: To account for the ratings that viewers append to each video, we convert each rating and its frequency into a score of +1 if the rating is positive (Funny, Beautiful, Ingenious, Courageous, Informative, Fascinating, Persuasive, Jaw-dropping, Inspiring) and a score of -1 if negative (Confusing, Unconvincing, Obnoxious,Long winded). Then we add up the score for each rating times its frequency to get the aggregate rating score. We note that there is a bias in the data for videos to be dis-proportionally positive, as such we introduce an average ratings that divides the aggregate ratings by the total count of all ratings.
- **Number of Related Talks**: We can treat the relationship between videos as a graph. A more popular video will be closer to the center of that graph, and popularity drops off from the center. Websites such as YouTube

will recommend videos that are popular so as to gain traffic on its website, so the number of related videos implies how many videos websites like YouTube will recommend viewers to watch. We can think of this graph as a social notwork, if many people are related to the center then that person is ”popular” in the conventional manner of social popularity, and individuals whom not many people know (ie are less ”popular”) will be on the fringes of that graph. We apply this logic to the videos by using how many talks a video is related to. The more related talks, the more popular the video.

Since all of these variables are on vastly different scales of magnitude, we normalize the variables and add them to create our composite popularity score.

3.2 Predictors

The predictors in this analysis are duration, number of speakers in the video, how old the video is, when the video was published, the ”sentiment” of the title and themes associated with the video, and the length of the title. Table 1 and 2 summarize the predictors.

- **Duration**: We include duration of videos as we assume that individuals are more likely to watch a shorter video that a longer video.
- **Number of Speaker**: We assume that with more speakers will increase the chance that a viewer can associate with the video and thus watch it.
- **Film Age**: We assume that more recently produced videos are more likely to be seen. For example, the data set contains videos from the 1990's to 2017. As the data set is from when the videos were published from 2006 to 2017, the views during this time will reflect a viewers bias to watch newer produced videos, that is to prefer videos from 2007 on wards over videos from the 1990s. We posit this assumption because individuals might discount older talks as outdated and thus not worth their time. Film Age is the videos age in days since being produced.
- **Video Age Group**: We add a categorical variable for when the video was published, labeled as 'old' for videos published prior to 2010 and young after 2010. We choose 2010 as that is when TED talk's underwent a sharp increase in the number of videos produced and would be a good divider between when TED was well know or not. We assume that viewers also consider when a video was published and have a preference for videos published more recently for the same reason as Film Age. This is different from when the variable Film Age, that is there is a difference between when a video was made and when it was published. A video made in 1990 but published in 2016 would be considered new, and viewers might consider the content relevant despite the date of production. However, this same video might still have a large Film Age coefficient as it has been in circulation for a year until 2017.
- **Title Sentiment**: As the title is the first thing a viewer will see, we assume that the title plays a crucial role in attracting views. To account for this we applied data clustering with K-Means with TF-IDF [1] on the titles to try to separate titles into three groups that might suggest the titles topic.
- **Title Length**: We assume that shorter titles, like shorter video lengths, will encourage views as a viewer can quickly understand the topic of the video rather than being forced to read a lengthy title which could be potentially cutoff, which could further disincentive a viewer.
- **Themes Label**: The data set provides a list of themes that the video is associated with. However, the videos will have several themes which might not be informative as TED talks has an incentive to apply as many themes to garnish views. To deal with this, we apply data clustering with K-Means with TF-IDF to determine the most relevant theme of the video. We include this predictor as some themes might have a larger following than other.

[1] K-Means with TF-IDF (term frequency–inverse document frequency) is a data clustering method that groups text by 'similarity'.

Variable	Levels	n	%	∑%
Title Label	future	2197	86.2	86.2
	life	85	3.3	89.5
	new	73	2.9	92.3
	world	195	7.7	100.0
	all	2550	100.0	
Video Theme	brain	147	5.8	5.8
	business	184	7.2	13.0
	culture	605	23.7	36.7
	design	329	12.9	49.6
	energy	64	2.5	52.1
	global	354	13.9	66.0
	health	192	7.5	73.5
	music	118	4.6	78.2
	science	346	13.6	91.7
	social	211	8.3	100.0
	all	2550	100.0	
Video Age Label	new	1711	67.1	67.1
	old	839	32.9	100.0
	all	2550	100.0	

Table 1

Variable	n	Min	q_1	$\widetilde{x}$	$\bar{x}$	q_3	Max	s	IQR	#NA
Average Views/Day	2550	17.0	311.0	724.0	1486.1	1752.8	28347.0	2148.2	1441.8	0
Video Duration	2550	135.0	577.0	848.0	826.5	1046.8	5256.0	374.0	469.8	0
Num. Speakers	2550	1.0	1.0	1.0	1.0	1.0	5.0	0.2	0.0	0
Film Age	2550	126.0	1177.2	2100.0	2230.9	2977.0	16667.0	1385.9	1799.8	0
Title Length	2550	1.0	5.0	6.0	6.2	8.0	16.0	2.3	3.0	0
Popularity	2550	0.6	1.6	1.8	1.7	1.9	3.2	0.2	0.2	0

Table 2: Numerical Response and Predictors

4 Statistical Analysis Without Mixed Models

In this analysis we use two models to determine the ”popularity” of TED talks using the two responses; average views per day and the composite popularity score. To predict the average views per day, which is a count variable, we use a Poisson regression. To predict the composite popularity score we use a linear regression.

4.1 Poisson Regression on Average Views Per Day

To predict the average views per day, which is a count variable, we use a Poisson regression with a log link.

$$\begin{aligned} log(\text{Video } i \text{ Average Num.Views/Day}) = &\text{ Video Duration}_i + \text{Num. Speakers}_i + \text{Film Age}_i + \text{Title Length}_i \\ &+ \text{Video Themes}_i + \text{Titles Content}_i + \text{Video Age Group}_i + \epsilon_i \\ i = &\{1, .., 2550\} \end{aligned} \tag{1}$$

This model indicates the log of the average number of views a video will obtain on any given day given the predictors.

4.2 Linear Regression on Composite Popularity Score

To predict the popularity, which is a normally distributed number, we use a Linear regression model.

$$\begin{aligned}\text{Video } i \text{ Popularity} &= \text{Video Duration}_i + \text{Num. Speakers}_i + \text{Film Age}_i + \text{Title Length}_i \\ &+ \text{Video Themes}_i + \text{Titles Content}_i + \text{Video Age Group}_i + \epsilon_i \\ i &= \{1, .., 2550\}\end{aligned} \tag{2}$$

This model indicates the popularity score a video will obtain given the predictors.

5 Statistical Analysis With Mixed Models

As these videos vary greatly in when they were published/created and their themes, we model how these differences might affect the response variables using Mixed Models for the previous regressions.

5.1 Time Variation

To determine whether the characteristics that predict popularity change over time, we define time into two groups, old (videos published prior to 2010) and new (after 2010). We would like to see how the relationship changes when TED talks had a surge in videos produced during 2010 that might be attributed to a change in an underlying demand for TED talk videos such as a new generation having access to the internet for example. We chose to include random intercepts and slopes. We use random intercepts because we believe that the viewers from the early 1990s are different from viewers from 2017. Indeed, viewers of TED talk videos in 1990 when personal computers where not readily available were most likely from a higher socio-demographic population than the average viewer in 2017 when the vast majority of individuals have personal computers. As such, 'old' and 'new' videos will have different intercepts (different average popularity or average views/day). Further, we include random slopes to account for potential differences in populations that might have different video viewing behaviors. Indeed, an individual in the early 2000s with a computer was most likely more educated (due to the general lack of widespread computers they would need it for very specific reasons like work/school but not for pleasure, generally speaking) than the average computer user in 2017 (where everyone owns a computer for both pleasure and work), and as such the former group might rate a longer video better than the latter group due to differences in attention span, for example.

5.1.1 Poisson Mixed Model Regression on Average Views Per Day: Time Variation

We apply Mixed Models to the Poisson regression to model the variation in time for the average views per day. $log(\text{Average Num.Views/Day}_{ij})$ denotes the log number of average views per day.

$$\begin{aligned}log(\text{Average Num.Views/Day}_{ij}) &= \beta_0 + b_{0j} + (\beta_1 + b_{1j})\text{Video Duration}_{ij} \\ &+ (\beta_2 + b_{2j})\text{Num. Speakers}_{ij} + (\beta_3 + b_{3j})\text{Film Age}_{ij} + (\beta_4 + b_{4j})\text{Title Length}_{ij} \\ &+ (\beta_5 + b_{5j})\text{Titles Content}_{ij} + \epsilon_{ij}\end{aligned} \tag{3}$$

$$i = \{1, .., n_j\}, j = \{old, new\}$$

In this model we have b's as the random slope/intercept for i observations from each time group j.

5.1.2 Linear Mixed Model Regression on Composite Popularity Score: Time Variation

We apply Mixed Models to the Linear regression to model the variation in time for the popularity score. Popularity_{ij} denotes the popularity score.

$$\begin{aligned}\text{Popularity}_{ij} &= \beta_0 + b_{0j} + (\beta_1 + b_{1j})\text{Video Duration}_{ij} \\ &+ (\beta_2 + b_{2j})\text{Num. Speakers}_{ij} + (\beta_3 + b_{3j})\text{Film Age}_{ij} + (\beta_4 + b_{4j})\text{Title Length}_{ij} \\ &+ (\beta_5 + b_{5j})\text{Titles Content}_{ij} + \epsilon_{ij}\end{aligned} \tag{4}$$

$$i = \{1, .., n_j\}, j = \{old, new\}$$

In this model we have b's as the random slope/intercept for i observations from each time group j.

5.2 Themes Variation

To determine whether the characteristics that predict popularity vary by the talks themes, we define themes into ten groups, as determined by the K-means clustering with TF-IDF, as: brain, business, culture, design, energy, global, health, music, science, social. As such we use a mixed model random intercepts and slopes to account for these differences. We use random intercepts because we believe that each theme attracts viewers from different populations and thus have different response averages. That is, viewers of science related videos are probably from a different population as those who watch culture related videos. Further, we include random slopes to account for potential differences in populations that might have different video viewing behaviors. Indeed, a computer science student that watches science related videos will most likely have a different attention span than the average viewer who will watch music related videos. As such, the former group might rate a longer video better than the latter group due to differences in attention span, to use that crude analogy twice.

5.2.1 Poisson Mixed Model Regression on Average Views Per Day: Theme Variation

We apply Mixed Models to the Poisson regression to model the variation in themes for the average views per day. $log(\text{Average Num.Views/Day}_{ij})$ denotes the log number of average views per day.

$$\begin{aligned} log(\text{Average Num.Views/Day}_{ij}) = \beta_0 + b_{0j} &+ (\beta_1 + b_{1j})\text{Video Duration}_{ij} \\ &+ (\beta_2 + b_{2j})\text{Num. Speakers}_{ij} + (\beta_3 + b_{3j})\text{Film Age}_{ij} + (\beta_4 + b_{4j})\text{Title Length}_{ij} \\ &+ (\beta_5 + b_{5j})\text{Titles Content}_{ij} + \epsilon_{ij} \end{aligned} \tag{5}$$

$$i = \{1, .., n_j\}, j = \{brain, business, culture, design, energy, global, health, music, science, social\}$$

In this model we have b's as the random slope/intercept for i observations from each theme j.

5.2.2 Linear Mixed Model Regression on Composite Popularity Score: Theme Variation

We apply Mixed Models to the Linear regression to model the variation in themes for the popularity score. Popularity_{ij} denotes the popularity score.

$$\begin{aligned} \text{Popularity}_{ij} = \beta_0 + b_{0j} &+ (\beta_1 + b_{1j})\text{Video Duration}_{ij} \\ &+ (\beta_2 + b_{2j})\text{Num. Speakers}_{ij} + (\beta_3 + b_{3j})\text{Film Age}_{ij} + (\beta_4 + b_{4j})\text{Title Length}_{ij} \\ &+ (\beta_5 + b_{5j})\text{Titles Content}_{ij} + \epsilon_{ij} \end{aligned} \tag{6}$$

$$i = \{1, .., n_j\}, j = \{brain, business, culture, design, energy, global, health, music, science, social\}$$

In this model we have b's as the random slope/intercept for i observations from each time group j.

6 Results without Mixed Models

6.1 Poisson & Linear Regression

We first look at the result from the Poisson (1) and Linear (2) regressions without mixed models. Table 3 summarizes the results. Lasso regression was used to select the most significant variables, however some variables were forced to be kept for the sake of comparing the two models.

Table 3: Poisson and Linear Regression

	Dependent variable:	
	avg_views_per_day	popularity
	Poisson	Linear
Duration	1.387*** (0.008)	0.132** (0.056)
Num. Speaker	-0.834*** (0.009)	-0.083(0.076)
Film Age in Days	-16.303*** (0.016)	-2.849*** (0.118)
Title Label: Life	0.149*** (0.002)	-0.001(0.022)
Title Label: New	-0.142*** (0.003)	-0.007(0.023)
Title Label: World	-0.099*** (0.002)	-0.044*** (0.015)
Title Length	0.053*** (0.004)	0.087*** (0.030)
Theme Label: Business	-0.135*** (0.002)	-0.002(0.022)
Theme Label: Culture	-0.314*** (0.002)	-0.043** (0.018)
Theme Label: Design	-0.537*** (0.002)	-0.057*** (0.019)
Theme Label: Energy	-0.589*** (0.004)	-0.082*** (0.029)
Theme Label: Global	-0.845*** (0.003)	-0.059*** (0.019)
Theme Label: Health	-0.692*** (0.003)	-0.004(0.021)
Theme Label: Music	-0.523*** (0.003)	-0.130*** (0.025)
Theme Label: Science	-0.669*** (0.002)	-0.030(0.019)
Theme Label: Social	-0.552*** (0.002)	-0.047** (0.021)
Video Age Group: Old	-1.841*** (0.006)	-0.317*** (0.026)
Film Age:Video Age Group: Old	13.460*** (0.031)	2.434*** (0.153)
Intercept	8.952*** (0.003)	2.038*** (0.025)
Observations	2,550	2,550
R^2		0.360
Adjusted R^2		0.356
Akaike Inf. Crit.	2,071,454.000	
Note:	*p<0.1; **p<0.05; ***p<0.01	

What is interesting is that the Poisson model kept everything as significant whereas the Linear model removed the title labels Life and New and the theme labels science and social. Further, both models agree on all signs and rankings in the case of categorical variables which adds to the robustness of these models ability to predict popularity.

Note, for this analysis we use terms such as increase or decrease for simplicity, however we note that the linear model increases/decreases the response by the amount of the reported value in Table **??**, whereas the poisson model increases/decreases the responses by exponentiating the reported value.

- **Duration**: Both models report an increase in the response, which is counter intuitive as we initially believed that viewers would be turned off from longer videos.
- **Number of Speaker**: Both models reported a decrease in the response, however the linear model did not consider it significant. This goes against our initial beliefs that more speakers would increase the likelihood of a viewer seeing a speaker they enjoy. We posit that this could be because more speakers could raise the chances that a viewer will dislike one of the speakers and not watch the video.
- **Film Age**: Both models report significant decreases in the response, which agrees with our initial hypothesis that viewers prefer videos that have been published more recently.
- **Video Age Group**: Both models report decreases in the response. This supports our initial hypothesis that viewers prefer videos that have been produced more recently.
- **Film Age:Video Age Group: Old**: We also model the interaction between these two terms to see how talks produced prior to 2010 but published after 2010 performs. Both models report an increase in the responses, which suggests that an 'old' video but published recently will perform just as well as a 'new' video published recently. This suggest that the age of the video in days since posted is a stronger predictor for success than the actual content of the talk.
- **Title Length**: Both models report increases in the response and are significant. This is surprising as we thought verbose titles would discourage viewers from watching the video.
- **Themes Label**: Both models report all the labels as significant and negative. Recall, each β_{theme} is the difference between the average response for that theme and the the average response for the base theme 'Brain'. Since all the themes are negative, this suggests that the best theme for talks is about the Brain. Further, as these coefficients act as rankings between the levels where the values closest to zero suggest higher popularity. Indeed, we see that 'Business' and 'Culture' are among the more popular themes while 'Global' and 'Health' are among the least popular. These results are true for both models
- **Title Label**: Using the same logic as above, we see that both models agree that titles regarding 'World' and 'New' are the least popular, that the base topic of 'Future' is the second most popular and 'Life' is the most popular. Both models only agree on the significance of 'World' which suggests that this topics popularity can be accurately predicted unlike the others.

7 Results with Mixed Models

7.1 Poisson & Linear Mixed-Effects

We now look at the result from the Poisson and Linear mixed models with random slope and intercept. Table 4 summarizes the results.

Table 4: Poisson and Normal Mixed-Effects with Random Slope and Intercept

	Dependent variable:			
	avg_views_per_day		popularity	
	Poisson Themes	Poisson Times	Normal Themes	Normal Times
Duration	1.976*** (0.490)	1.976*** (0.490)	0.399*** (0.136)	0.382*** (0.128)
Title Length	0.155(0.169)	0.155(0.169)	0.178*** (0.050)	0.197*** (0.030)
Title Label: Life	0.064(0.093)	0.064(0.093)	0.023(0.032)	0.006(0.022)
Title Label: New	-0.166(0.147)	-0.166(0.147)	-0.0002(0.027)	-0.003(0.024)
Title Label: World	-0.129** (0.054)	-0.129** (0.054)	-0.048** (0.023)	-0.043*** (0.015)
Num. Speaker	-0.689(0.755)	-0.689(0.755)	-0.047(0.124)	-0.048(0.123)

Table 4: Poisson and Normal Mixed-Effects with Random Slope and Intercept

Film Age	-12.327*** (0.698)	-12.327*** (0.698)	-1.330*** (0.145)	-1.309*** (0.131)
Intercept	8.105*** (0.104)	8.105*** (0.104)	1.797*** (0.035)	1.791*** (0.034)
Observations	2,550	2,550	2,550	2,550
Log Likelihood	-1,075,431.000	-1,075,431.000	436.955	429.853
Akaike Inf. Crit.	2,150,950.000	2,150,950.000	-783.910	-821.705
Bayesian Inf. Crit.	2,151,207.000	2,151,207.000	-520.937	-710.672
		Random Effects Variance:		
Intercept	0.10631	1.8097	0.0093	0.0093
Duration	2.22376	0.04407	0.1414	0.1413
Num. Speaker	4.80851	0.106	0.0833	0.0833
Film Age	4.09735	15.22	0.1711	0.1711
Title Length	0.27191	0.728	0.0144	0.0141
Title Label: Life	0.08319	1.063	0.0043	0.0043
Title Label: New	0.2113	0.79156	0.00126	0.00124
Title Label: World	0.0276	0.3207	0.00244	0.0401
Observations	2,550	2,550	2,550	2,550
Log Likelihood	-1,075,431.000	-1,075,431.000	436.955	429.853
Akaike Inf. Crit.	2,150,950.000	2,150,950.000	-783.910	-821.705
Bayesian Inf. Crit.	2,151,207.000	2,151,207.000	-520.937	-710.672
Note:				*p<0.1; **p<0.05; ***p<0.01

Since mixed models will produce the same coefficients as its non mixed counterpart we skip the description of the coefficients. Further, all the models have very similar coefficients as their non mixed counterparts except some are slightly different which could be because of convergence issues with the lmer package. The interpretations and findings from the previous section hold the same.

7.1.1 Model Selection

We now test the random intercepts and random slopes. That is, we test whether there exists difference in the average responses across time and themes, and whether there exists differences in how the predictors effect the response across time and themes. To compare a generalized mixed model (GLMM) without a random component with a GLMM with a random component in R, we had to use the GLM package for the nested model and LMER4 for the full model as LMER4 does not allow models to not have a random component. Evidently, to compare the nested and full model with random intercepts, we used AIC as our criterion because the ANOVA function in R is not able to compare models from the LMER4 and GLM packages. Table 5 summarizes the results of testing the random intercept.

- **Linear Model with Random Intercept for Themes**: The AIC for the mixed model with random intercept is lower than the null, so it is preferred. This suggest that there exists differences in the average popularity score across themes.
- **Linear Model with Random Intercept for Time**: The AIC for the mixed model with random intercept is lower than the null, so it is preferred. This suggest that there exists differences in the average popularity score across time, that is for videos published before and after 2010.
- **Poisson Model with Random Intercept for Themes**: The AIC for the mixed model with random intercept is lower than the null, so it is preferred. This suggest that there exists differences in the average avg. views per day across themes.

- **Poisson Model with Random Intercept for Time**: The AIC for the mixed model with random intercept is lower than the null, so it is preferred. This suggest that there exists differences in the average avg. views per day across time, that is for videos published before and after 2010.

Table 5: Testing Random Intercept

	df	AIC
Linear Popularity Themes null	9.00	-810.37
Linear Popularity Themes full	10.00	-789.73
Linear Popularity Times null	9.00	-810.37
Linear Popularity Times full	10.00	-767.85
Poisson Avg. Views/day Themes null	8.00	2456258.87
Poisson Avg. Views/day Themes full	9.00	2262017.56
Poisson Avg. Views/day Times null	8.00	2456258.87
Poisson Avg. Views/day Times full	9.00	2418695.71

Similarly, we test the random slopes. We test these random slopes using χ^2 model selection. Table 6. summarizes the results.

- **Linear Model with Random Slope for Themes**: The p-value is very small, so the model without random slope is not an adequate simplification of the full model; the preferred model includes the random intercept. This suggest that there exists differences in how the predictors effect popularity score across themes.
- **Linear Model with Random Slope for Time**: The p-value is very small, so the model without random slope is not an adequate simplification of the full model; the preferred model includes the random intercept. This suggest that there exists differences in how the predictors effect popularity score across time, that is for videos published before and after 2010.
- **Poisson Model with Random Slope for Themes**: The p-value is very small, so the model without random slope is not an adequate simplification of the full model; the preferred model includes the random intercept. This suggest that there exists differences in how the predictors effect avg. views per day across themes.
- **Poisson Model with Random Slope for Time**: The p-value is very small, so the model without random slope is not an adequate simplification of the full model; the preferred model includes the random intercept. This suggest that there exists differences in how the predictors effect avg. views per day across time, that is for videos published before and after 2010.

Table 6: Testing Random Slopes

	Df	Chisq	Chi Df	Pr(>Chisq)
Linear Popularity Themes simple	10			
Linear Popularity Themes full	45	59.91	35	0.0055
Linear Popularity Times simple	10			
Linear Popularity Times full	19	240.34	9	0.0000
Poisson Avg. Views/day Themes simple	9			
Poisson Avg. Views/day Themes full	44	111138.01	35	0.0000
Poisson Avg. Views/day Times simple	9			
Poisson Avg. Views/day Times full	44	155142.81	35	0.0000

Clearly, the best models include both random intercepts and slopes for both Poisson and Linear models. Which suggests that a lot of the variation in the data can be explained by timing of when a video is published and its theme rather than the actual form/content. The main predictors in these models are the talks duration, talks film age in days since published, and the talks age group (when the talk was published) are the strongest predictors of popularity. This suggests that the best way to create a popular video is by frequently publishing long videos, regardless of the content. Popularity seems to be a function of quantity not quality.

Due to computational restrictions we were unable to run mixed models with both time and themes as random components.

8 Conclusion

In this analysis we used web scraped data from the TED talks website to determine what predicts talks popularity. To measure popularity we used average views per day and a composite score that encompasses several intuitive measures of popularity. We used Linear and Poisson regression with Lasso penalization and Linear and Poisson Mixed Models with random intercepts and slopes to exploit variation in talks across time and themes. We found that the majority of variation in the data can be explained by variation across time and themes. Further, we found that the strongest predictors of popularity, in all models, was a talks duration and how long it has been uploaded for.

Future research will benefit from using the transcripts of the talks as a measure of the talks content. Further, due to computational restrictions we were unable to run a model with both time and themes as random components. Future research will benefit from running these models to see which component explains the data the most.

References

[1] Mi-Kyung Song, Feng-Chang Lin, Sandra E Ward, and Jason P Fine. Composite variables: when and how. *Nursing research*, 62(1):45, 2013.