Predicting the Popularity of TED Talks Using Composite Measures of Popularity and Mixed Models

Marcos Cardenas Zelaya

March 10, 2020

Abstract

Objectives: Determine the relationship between osteoarthritis (OA) and cardiovascular disease using Canadian survey data.

Design: Logistic Mixed-Models Regression is used to determine the odds ratio between OA and heart disease.

Data: Canadian Community Health Survey (CCHS) from 2000 to 2005.

Participants: Adult participants aged 20-64 in the CCHS cycles 1.1, 2.1 and 3.1 were included. CCHS dataset includes nationally representative data on heart disease and other health determinants. All observations (responses and predictors) are self-reported from 10 provinces and 3 territories. We have selected 200,478 observations. Observations are not identifiable between cycles.

Predictors and Response: Cardiovascular disease is the response. The main predictor is OA after adjusting for socio-demographic factors, access to a doctor, obesity, physical activity, smoking status, drinking status, diabetes and hypertension.

Results: There is no evidence to suggest that OA is associated with heart disease. There is also little evidence to suggest that the association between OA and heart disease vary across gender, marital status, region and recency of immigration.

Conclusion: Accounting for demographics, OA is not associated with heart disease. Due to computational restrictions, more research is required to accurately asses the relationship between OA and heart disease.

1 Introduction

2 Data

We used descriptive data on TED Talk videos from 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 wold 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 an 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.
- **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 for 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 rough 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 **??** 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 this will increase the chance that a viewer is can associate with the video and thus watch it.
- **Film Age**: We assume that more recent 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 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.

- **Video Age Group**: We add a categorical variable for when the video was published, labeled as 'old' for videos 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. This is different from when a video was made. 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.
- **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 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 topic of the video rather than being forced to read a lengthy title which could 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. For example themes regarding science might garnish more views than themes regarding plumbing.

Variable	**Levels**	**n**	%	$\sum$%
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.

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

$$i = \{1, .., 2550\}$$

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.

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

$$i = \{1, .., 2550\}$$

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. That is, viewers of science related videos are probably from a different population as those who watch culture related videos. As such, we assume each theme will attract different populations and thus videos from each theme 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 computer science student that watches science related videos will most likely have a different attention span than a 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	1.732*** (0.008)	0.258*** (0.058)
Num. Speaker	-0.823*** (0.009)	-0.035(0.079)
Film Age in Days	-14.052*** (0.014)	-1.433*** (0.081)
Title Label: Life	0.158*** (0.002)	0.005(0.023)
Title Label: New	-0.144*** (0.003)	-0.002(0.024)
Title Label: World	-0.096*** (0.002)	-0.043*** (0.015)
Title Length	0.196*** (0.004)	0.216*** (0.031)
Theme Label: Business	-0.132*** (0.002)	-0.006(0.023)
Theme Label: Culture	-0.306*** (0.002)	-0.041** (0.019)
Theme Label: Design	-0.511*** (0.002)	-0.052** (0.020)
Theme Label: Energy	-0.579*** (0.004)	-0.082*** (0.031)
Theme Label: Global	-0.888*** (0.003)	-0.079*** (0.020)
Theme Label: Health	-0.681*** (0.003)	-0.002(0.022)
Theme Label: Music	-0.477*** (0.003)	-0.118*** (0.026)
Theme Label: Science	-0.655*** (0.002)	-0.030(0.020)
Theme Label: Social	-0.493*** (0.002)	-0.013(0.022)
Video Age Group: Old	0.514*** (0.002)	0.035*** (0.013)
Intercept	8.706*** (0.003)	1.852*** (0.024)
Observations	2,550	2,550

Table 3: Poisson and Linear Regression

R^2		0.296
Adjusted R^2		0.291
Akaike Inf. Crit.	2,217,004.000	
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 tittle labels Life and New and the theme labels

7 Conclusion

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.