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

**Marcos Cardenas Zelaya**

March 29, 2020

## 1 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.

### 1.1 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.

### 1.1.1 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.

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

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

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

### 1.1.2 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. $\text{Popularity}_{ij}$ denotes the popularity score.

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

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

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

## 1.2 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.

### 1.2.1 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.

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

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

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

### 1.2.2 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. $\text{Popularity}_{ij}$ denotes the popularity score.

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

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

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

$$\begin{aligned}\text{y}_{ij} &= \beta_0 + b_{0j} + \beta_1\text{spiciness}_{ij} + \beta_2\text{meatamount}_{ij} \\ &+ (\beta_3 + b_{1j})\text{taco type}_{ij} + \varepsilon_{ij}\end{aligned} \tag{5}$$

$$i = \{1, ..., n_j\}, j = \{tacobell, ..., chaleteco, pueblito\}$$

## References