Star Trek: The Analysis

Duke University IDS702: Final Project

Satvik Kishore

December 12, 2021

https://github.com/satvikk/StarTrek_StatsFinalProject

Summary

Star Trek is a Science Fiction media franchise spanning several TV shows, movies, books, and video games. This analysis focused on four Star Trek TV shows that ran in sequence from 1987 through 2005, with significant overlaps in the years they each ran. The questions of interest are:

- Which characters have the greatest influence on the ratings of episodes, either positive or negative?
- Are their differences in quality across the directors?
- Are their differences in quality across the four shows and their constituent seasons?

To answer these questions, I use the IMDb ratings of the episodes. I model the data using hierarchical linear regressions and answer the questions using the estimates from this model. The most important results were that there are significant differences in quality across the seasons of the shows. Some of the characters are found to have strong associations between their screentime and corresponding episode ratings. Screentime of Sisko from Star Trek: Deep Space 9 and Paris from Star Trek: Voyager have the strongest positive impact on episode ratings whereas screentime of Dr. Crusher, Troi, and Wesley from Star Trek: The Next Generation, and Dax from Star Trek; Deep Space 9 have the strongest negative impact on the ratings.

Introduction

Star Trek is one of the most successful science fiction media franchises and was an instant hit during its inception in the 1960s. The series continues to be relevant, with most of its TV shows being available on Over-the-Top services like Netflix. As of December 2021, there are three TV shows running, with another one set to begin in 2022. In this analysis, I have chosen the shows The Next Generation, Deep Space 9, Voyager, and Enterprise. These four shows ran from 1987 through 2005 and constitute the largest block of content produced by the franchise. I have omitted The Original Series as well as the newer shows because of the separation in time between them. The four shows in consideration here have a similar system, with a consistent set of personnel behind their production. The shows are mostly episodic in format, i.e. each episode is self contained within its plot, having an introduction, body, and conclusion. The episodes are not completely independent however, with plot elements sometimes having an effect on the larger story. There are a few multi-parter episodes as well. The shows each have a set of core characters that have varying degrees of screentime in each episode. In addition, there are minor recurring characters and usually a few one-time episodes. Among the community, there are differing opinions on the qualities of individual core characters, but there are a clear few fan-favorites and also some that are widely disliked. My first question pertains to these perceptions, on as to how character screentimes are associated with episode ratings. As is the case with TV shows, there are many directors, and the quality of directors is expected to have a strong effect on the quality of the episodes. Therefore, another research question of mine is: does the quality of directors differ, and if yes, who are the best (or worst). My third question is on differences in quality between the shows, and their constituent seasons.

Data

The two sources of data are the IMDb datasets provided on datasets.imdbws.com (collected on Nov 23, 2021), and script data provided on www.chakoteya.net/StarTrek/index.html. The IMDb data is available in a relational schema form. It is very large and has most of IMDb data contained within it. I filtered and wrangled to select only the four shows in consideration, with the variables: director name, names of writers, episode number, season number, and the IMDb rating. A few episodes have multiple directors. Combination of directors are considered to be a different director. Most episodes have multiple writers, and this variable is dropped from further analysis. The script data is provided as a json, formatted as Show -> episode -> character -> all lines. For each character, within each show, I computed total number of words spoken by the character in the episode, and divide it by the total number of words spoken in the entire episode. This results in a proxy for the screentime of each character in each episode as a percentage with the value being between 0 and 100. I only select the main characters from this transformed data and merge it with the IMDb data. There were a few issues with matching episodes between these two datasets as the scripts data sometimes assumes multi-part episodes as single episodes. These episodes are split before merging, and the final data assumes screentime values for each episode in these multi-part episodes to be the same.

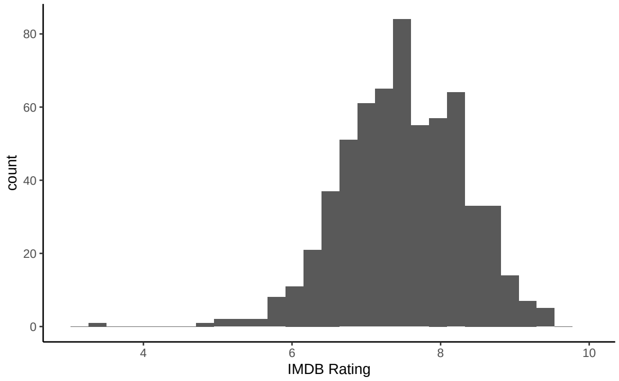
Exploratory Data Analysis:

Show	Abbrv.	n_seasons	n_episodes	n_directors	n_characters
The Next Generation	TNG	7	176	38	8
Deep Space 9	DS9	7	173	33	9
Voyager	VOY	7	168	33	10
Enterprise	ENT	4	97	20	7
Total		25	614	73	33

Table 1: Data Summary

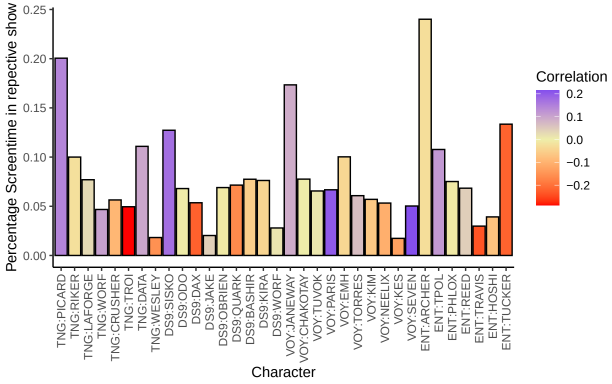
Histogram of Ratings across episodes

Figure 1



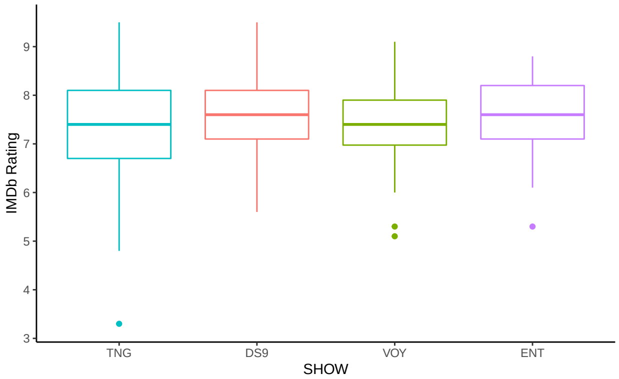
Correlation of IMDb Rating with percentage sceentime of characters

Figure 2



Boxplot of ratings by shows

Figure 3



Boxplot of ratings by seasons in TNG

Figure 4

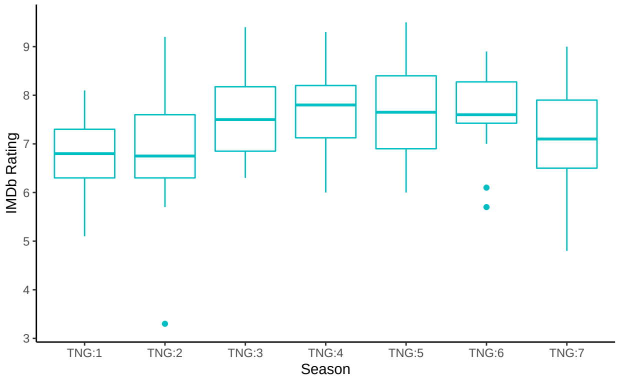


Table 1 provides information on the structure of the data, including counts of variables. Figure 1 illustrates the distribution of the `IMDb ratings`. The distribution appears to be roughly normal, permitting further analysis without the need for variable transformations. Figure 2 illustrates the correlation of `character-screentime` with the `IMDb ratings`, along with what percentage of screentime the each character occupies in their respective show. There are strong negative correlations for "TNG: Troi" and "DS9: Dax", and strong positive correlations for characters like "DS9: Sisko" and "TNG: Picard". Figure 3 illustrates distributions of ratings within each show. The shows appear to have similar ratings, with the median of TNG being a little lower than the medians of the other shows. Figure 4 shows the boxplots of ratings across seasons within TNG. We observe significant variations across seasons, with the later seasons being somewhat better rated. Similar trends can be observed for other shows as well (not illustrated).

Modeling

I use a linear hierarchical model to answer the questions of interest. The equation of the final model is:

$$y_i = \beta_0 + \gamma_{0,j_i} + \gamma_{1,k_i} + \sum_{p=1}^{33} \beta_p x_{p,i} + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

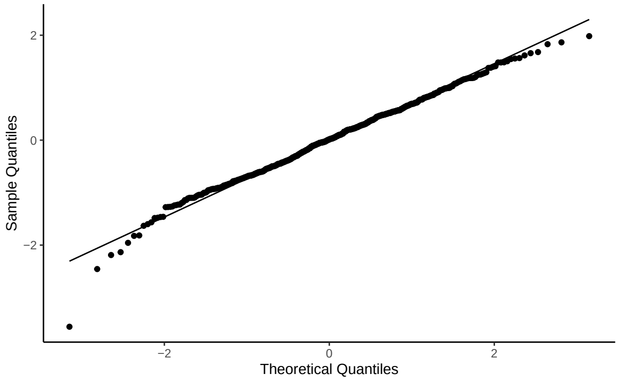
$$(\gamma_{0j}, \gamma_{1k}) \sim \mathcal{N}_2(\mathbf{0}, \Sigma)$$

where i is the observation index, y is the IMDb rating, β_0 is the grand intercept of the model, γ_{0,j_i} is the random intercept effect from the `show:season`, j_i is the show:season for episode i, γ_{1,k_i} is the random intercept effect from the `director`, k_i is the director for episode i, p is index for one of the 33 characters, β_p is the linear coefficient for the `screentime` of character p, $x_{p,i}$ is the `screentime` value of character p in episode i, ϵ_i is the residual term, σ^2 is the variance of the residual, and Σ is the covariance matrix for the three residual terms. I don not employ a random slopes model as the current model already has quite a large number of variables, especially for the given small data, and introduction of more random effects make the model too complex while damaging interpretability. I compare this model to a few others using Analysis of Variance tests. The model significantly better than the model with only director and season:show variables, indicating that the cast variables are important. It is also significantly better than a model without any of the random effects, indicating that the heirarchy levels are important as well. It is not significantly different than the model which includes position of episode within season as a fixed effect variable, and hence that variable is dropped from the analysis. It is also not significantly different from the model that has `show` as another random effect, and hence that hierarchy level is also dropped.

Model Assesment

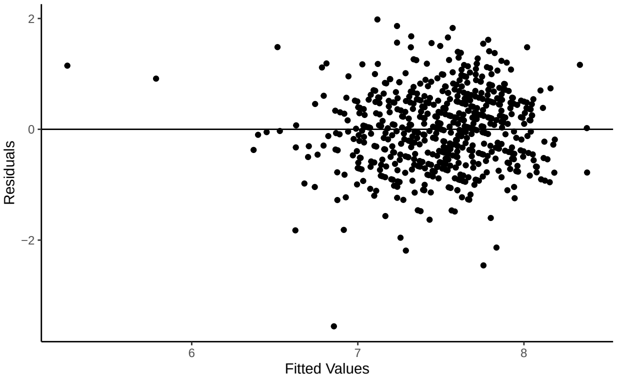
QQ plot of Model Residuals

Figure 5



Fitted vs Residuals

Figure 6



The final model satisfies the basic model assumptions. The error residuals are independent of each other due to the episodic nature of the shows, and the within season interdependence being taken care of through inclusion of the `show:season` hierarchy. The scatter plot of the residuals also appears sufficiently random (not included in report). The residuals are normally distributed as evidenced by the QQ plot in Figure 5.

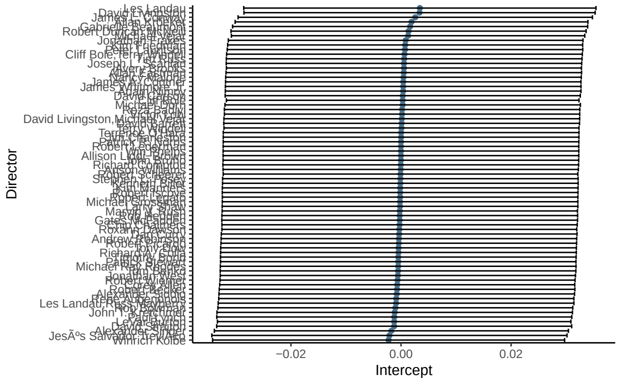
The Fitted vs Residuals plot in Figure 6 also suggests that the homoskedasticity and linearity assumptions are satisfied.

Results

In the random effects, the `show:season` level captures 11.55% of the variation. The `director` level captures 0.04% of the variation, strongly indicating that the `show:season` has a strong association with the ratings of its constituent episodes, whereas there is insufficient evidence to conclude the same for `director`. This is also evidenced in figures 7 and 8, where we can see that there is a large overlap in the confidence intervals of the effects of directors, but there are show:seasons that can be said to be better than other show:seasons with strong confidence. The confidence intervals in this discussion are all 95% confidence intervals. TNG:5, TNG:6, and ENT:4 are clearly far better than the likes of TNG:1 and TNG:2. There are seven different show:seasons that have their confidence intervals beyond the third worst season, i.e. DS9:1. The spread of TNG across both ends of the chart partially explains why using `show` as an hierarchical variable does not improve the model, as there is large variation in seasons within shows. The values of these random effects can be interpreted as, for example DS9:1 with a coefficient of -0.37 means that holding other variables constant, episodes in the first season of Deep Space 9 are on average rated 0.37 points less than the overall mean.

Random Effects for Directors

Figure 7



Random Effects for Show:Seasons

Figure 8

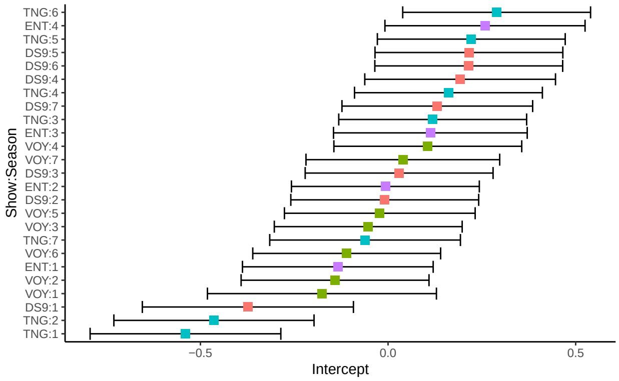
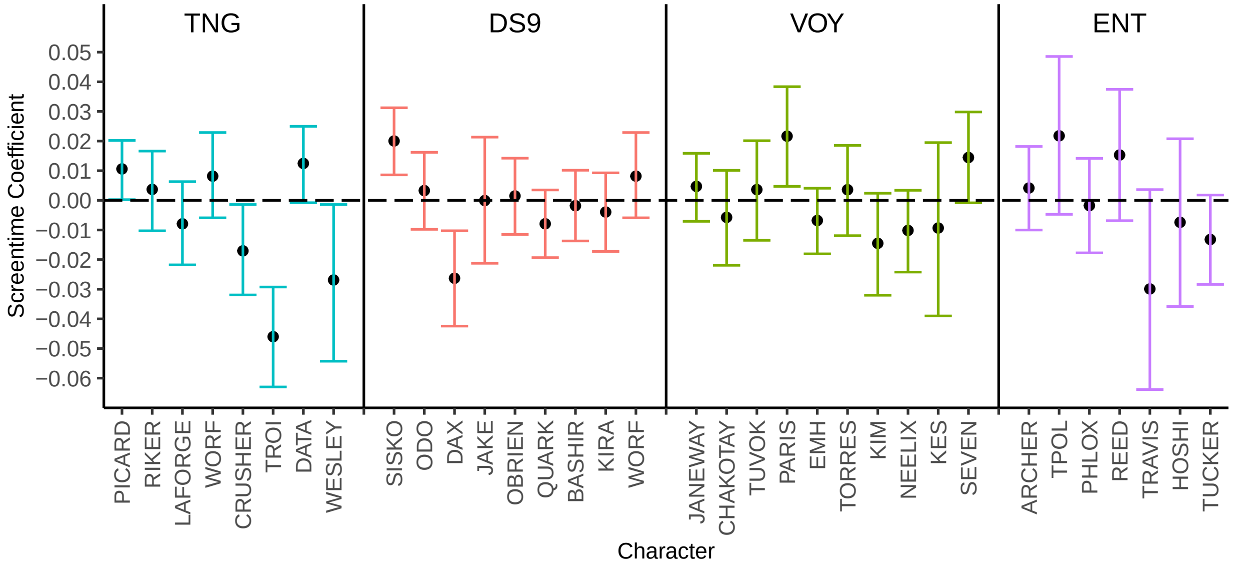


Figure 9 illustrates the estimates of the fixed effects, where we say that the effects that have their confidence interval to be beyond zero are significant at the 95% confidence level. Our fixed effects are the coefficients for the `character-screentime` variables. These coefficients can be interpreted as, for example for VOY:PARIS with a coefficient of 0.022, we can say that controlling for other variables, a one percentage point increase in screentime for PARIS and a one percentage point decrease in screentime for "other" characters is associated with 0.022 point increase in IMDb rating. The "other" character refers to any character not included in the model. For most characters, we find no evidence that their `screentime` is associated with the IMDb rating. But some of them are most certainly significant. TNG:TROI stands out as the worst character, and this result agrees with the general consensus in the community that despite having constant presence in the show, the character is poorly written and not developed well. Within TNG, we also see CRUSHER have a bad rating. This is interesting as she was the only other main female character other than TROI, and her poor rating could be indicative of the fact that the show runners were poor at developing their female characters. The other significant character within TNG was WESLEY. This character is famously disliked, and this dislike has been referenced in other TV shows as well. The show runners did scrap this character after the fourth season. The other significant characters are DS9:SISKO, DS9:DAX, and VOY:PARIS. There are no significant characters in ENT. This is likely due to the lower number of episodes in this show, resulting in less data.

Estimates of screentime coefficients for each character

Figure 9



Conclusion

Star Trek IMDb data from four shows is analyzed in context of three questions that we set out to answer. The method used to analyse the characters was the most interesting and can be easily extended to other TV shows as well. The concepts used here can be used by current showrunners to analyze their episodes and characters, and perhaps plan future episodes accordingly. The results indicate that:

- There are strong differences across show/season combinations, with at least seven individual seasons significantly found to be better than the third worst season over all four shows.
- The characters with strongest positive influence are Sisko from Deep Space 9 and Paris from Voyager. The characters with strong negative influence are Dr. Crusher, Troi, and Wesley from The Next Generation, and Dax from Deep Space 9.
- No evidence for differences in quality across directors is observed.

Limitations and Future Work

- While the hierarchical structure of the model helps in taking care of dependence in residuals within seasons, it is still possible that a few of the multi-part episodes have non independent residuals.
- The screentime is calculated as proportion of words spoken by characters. The source data is not fully clean as a few whitespace issues may have fused words together and reduced counts. Incidence of this issue is not very common and is completely at random, and thus should only have a minor influence on the results.
- IMDb calculates ratings as a weighted mean from individual user ratings, with the weight calculation algorithm being a secret. While the weighting scheme is in place to prevent manipulation, it leaves an uncertainty as to exactly what our outcome variable is.
- Characters change and evolve through a show's natural progression, and thus their effect on episode ratings is likely to be variable between seasons. This effect can be modeled using a random slopes model however the multifold increase in number of variables makes the modeling difficult. Thus we are forced to assume that each character has a uniform linear effect on episode ratings.
- The large number of directors, with small number of episodes per director makes it difficult to draw any conclusions about their quality from this data.
- Future work: We can also look at episode writers and variation in their quality. This will however be a challenge due to variable number of writers across episodes, and almost each episode being worked on by a unique combination of them.