Predicting Bechdel Test Results through Statistical Modeling

Hailey Quintavalle

Abstract

The Bechdel Test is a simple measurement designed to analyze the representation of women in film. To pass, a movie must have two female characters who have a conversation that is not about a man. The present research aims to use genre, release year, movie budget, user ratings, and critics rating scores to predict the probability a movie will pass the test. Logistic regression analysis reveals more recent movies are predicted to have a higher probability of passing the test. For most years, genres such as Romance and Comedy are predicted to have a higher probability of passing, while genres such as Action, Sport, War, and Western are predicted to have a lower probability of passing the test. Although the Bechdel Test has its flaws, the test is a useful metric to bring attention to the roles women hold in film.

Introduction

If you think of your all time favorite movie, try to think about how many of the main characters were women. In 2023, 35% of speaking roles in movies belonged to women, and only 28% of the top grossing films contained female protagonists (Lauzen, 2024). Just one decade before, in 2013, women occupied 30% of all speaking roles and just 15% of protagonists were women (Lauzen, 2014). This is a small change for a span of 10 years, showcasing the underrepresentation of women in film, and the slow steps towards progress the film industry seems to be taking. As we examine the portrayal of women's roles in film, it is useful to explore various criteria used to measure their representation. One notable measure is the Bechdel Test, unintentionally introduced by American cartoonist Alison Bechdel in 1986, and the primary measure of women in film that is the focus of the present analysis.

The Bechdel Test originated from a comic strip titled "The Rule", a part of her comic *Dykes To Watch Out For*. The comic depicts two women discussing certain criteria necessary for them to watch a movie. The criteria for a movie included: two named women, who talked to each other, and held a conversation with each other that was not about a man. In the early 2000s, the test quickly gained popularity online and has since been used as a common tool for analyzing the role of women in film.

The test has only three basic requirements, meaning that a film can pass with just one line. Because of the simplicity of this tool, many have come forward with fair criticisms of the test. For example, the test does not take into account demographic factors, such as age, the voices of women of color, or those who do not speak English as their first language (O'Meara, 2016). Other flaws of the test include the oversight of conversations that are not directly about a man, but may be indirectly about them, where the conclusion of the test are unclear.

Regardless of its flaws, the Bechdel Test is a straightforward and easy tool that allows people to quickly make base assumptions about the presence of women in film. For this reason, this analysis will be examining the characteristics of thousands of movies in an attempt to use logistic regression analysis to model the likelihood of passing the Bechdel Test.

Read the comic here.

Data

The data used for the present analysis comes from a combination of multiple online sources. A data set containing 10,183 movie titles was available through IMDb Non-Commercial Datasets. Release dates for these movies range from 1874 to 2023. Variables included in this data set can be seen in the preview of the data below.

IMDb Non-Commercial Data set:

```r
basics_movies |>
  select(tconst, titleType, primaryTitle, originalTitle,
         isAdult, startYear, runtimeMinutes, genres) |>
  head() ## |> pander::pander()
```

```
# A tibble: 6 x 8
  tconst   titleType primaryTitle originalTitle isAdult startYear runtimeMinutes
  <chr>    <chr>     <chr>        <chr>           <dbl>     <dbl> <chr>
1 tt31557~ short     Passage de ~ Passage de V~       0      1874 1
2 tt14495~ short     La Rosace M~ La Rosace Ma~       0      1877 1
3 tt22214~ short     Sallie Gard~ Sallie Gardn~       0      1878 1
4 tt12592~ short     Le singe mu~ Le singe mus~       0      1878 1
5 tt78164~ short     Athlete Swi~ Athlete Swin~       0      1881 1
6 tt54597~ short     Buffalo Run~ Buffalo Runn~       0      1883 1
# i 1 more variable: genres <chr>
```

```r
## test_df <- read_tsv(here::here("data/title.basics.tsv")) |>    # data downloaded from fro
##   mutate(imdbid = parse_number(tconst))
```

Other data that was used in this analysis comes from the Bechdel Test Movie List, where users can submit movies with their Bechdel test rating through their online platform. The data set pulled from this website contains 10,251 movies with release dates ranging from 1874 to 2024. For the rating variable, a movie is given a rating from one to three, directly corresponding with the number of requirements of the Bechdel test that it passes. A preview of this data can be shown below.

```r
movies_api <- read_csv("data/movies_api.csv")
movies_api |> head()
```

```
# A tibble: 6 x 5
     id title                       imdbid    year rating
  <dbl> <chr>                       <chr>    <dbl>  <dbl>
1  9602 Passage de Venus            3155794   1874      0
2  9804 La Rosace Magique           14495706  1877      0
3  9603 Sallie Gardner at a Gallop  2221420   1878      0
4  9806 Le singe musicien           12592084  1878      0
5  9816 Athlete Swinging a Pick     7816420   1881      0
6  9831 Buffalo Running             5459794   1883      0
```

The final data set that was used in this analysis comes from the TidyTuesday social data project through GitHub. This data set contains 1,794 movies released from 1970 up to 2013.

```
movies <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data

movies |> select(-error) |> head()
```

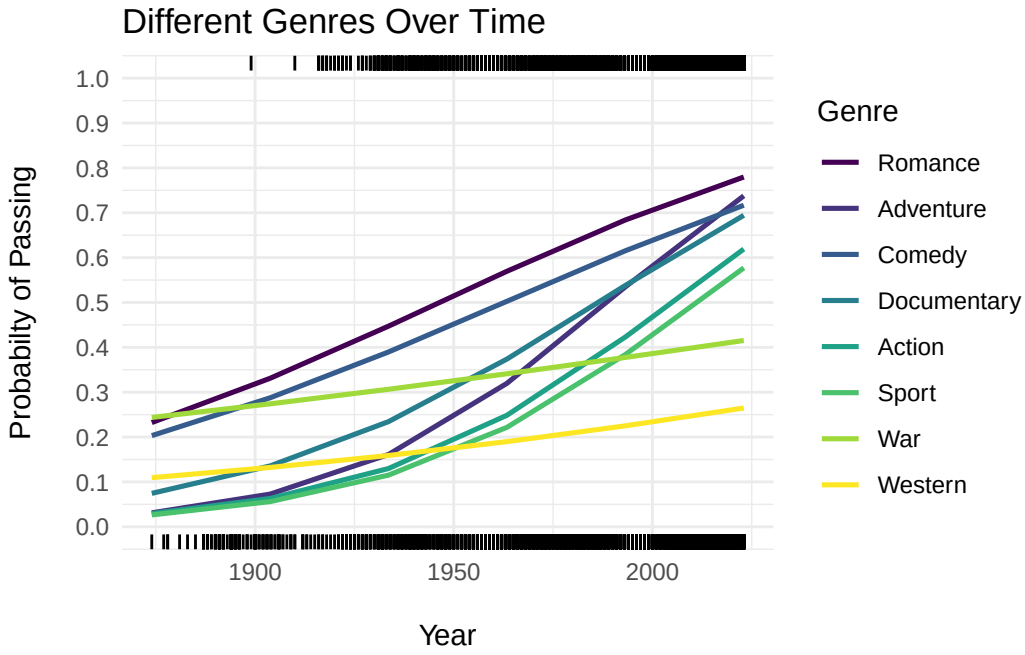
```
# A tibble: 6 x 33
   year imdb      title   test  clean_test binary budget domgross intgross code
  <dbl> <chr>     <chr>   <chr> <chr>      <chr>   <dbl> <chr>    <chr>    <chr>
1  2013 tt1711425 21 &am~ nota~ notalk     FAIL   1.3 e7 25682380 42195766 2013~
2  2012 tt1343727 Dredd ~ ok-d~ ok         PASS   4.50e7 13414714 40868994 2012~
3  2013 tt2024544 12 Yea~ nota~ notalk     FAIL   2   e7 53107035 1586070~ 2013~
4  2013 tt1272878 2 Guns  nota~ notalk     FAIL   6.1 e7 75612460 1324930~ 2013~
5  2013 tt0453562 42      men   men        FAIL   4   e7 95020213 95020213 2013~
6  2013 tt1335975 47 Ron~ men   men        FAIL   2.25e8 38362475 1458038~ 2013~
# i 23 more variables: budget_2013 <dbl>, domgross_2013 <chr>,
#   intgross_2013 <chr>, period_code <dbl>, decade_code <dbl>, imdb_id <chr>,
#   plot <chr>, rated <chr>, response <lgl>, language <chr>, country <chr>,
#   writer <chr>, metascore <dbl>, imdb_rating <dbl>, director <chr>,
#   released <chr>, actors <chr>, genre <chr>, awards <chr>, runtime <chr>,
#   type <chr>, poster <chr>, imdb_votes <dbl>
```

Modeling

Full model with all genres and all interactions:

[insert model] ??why wont latex work here??

```
list3 |> filter(values == 1, genre %in%
                  c("Sport", "Comedy", "Romance", "Adventure", "Action", "Documentary", "War
  mutate(genre = fct_reorder2(genre, .x = year, .y = .pi)) |>
  ggplot(aes(x = year, y = .pi)) +
  geom_line(aes(color = genre), linewidth = 0.9) +
  geom_rug(data = movies_pass, aes(x = year, y = as.numeric(binary)), sides = "t") +  ## alp
  geom_rug(data = movies_fail, aes(x = year, y = as.numeric(binary)), sides = "b") +
  scale_y_continuous(breaks = seq(0, 1, by = 0.1), limits = c(0, 1)) +
  theme_minimal(base_size = 11) +
  labs(x = "\nYear",
       y = "Probabilty of Passing\n",
       color = "Genre",
       title = "Different Genres Over Time") +
  scale_color_viridis_d()
```



Conclusion

References

1. Alison Bechdel, 1986. *Dykes to watch out for*. Firebrand Books.
2. Lauzen, M. (2024). It's a Man's (Celluloid) World: Portrayals of female characters in the top grossing U.S. Films of 2023. *Center for the Study of Women in Television & Film*. https://womenintvfilm.sdsu.edu/its-a-mans-celluloid-world-portrayals-of-female-characters-in-the-top-grossing-u-s-films-of-2023/
3. Lauzen, M. (2014). It's a Man's (Celluloid) World: On-Screen Representations of Female Characters in the Top 100 Films of 2013. *Center for the Study of Women in Television & Film*. https://womenintvfilm.sdsu.edu/files/2013_It's_a_Man's_World_Report.pdf