Can We Predict if a Movie Will Pass the Bechdel Test?

Jonathan Barnes

Joe Cyriac

Audrey Jenkins

Tara Sothy

Group A, Fall 2021

Abstract:

The goal of this project was to use data from Fivethirtyeight (bechdeltest.com) and Kaggle (IMDb), which are linked below. To create logistic models and make predictions if a movie passes the Bechdel test based on several possible predictors, such as average ratings from men, average ratings from women and if the IMDb description contains he/him pronouns or she/her pronouns. In class, we had the opportunity to explore several data sets and examine different relationships within. A topic we all found interesting and wanted to look more into was the performance of films and how different factors impact whether or not a movie might pass the Bechdel test. For this project, we decided to explore how several characteristics of a movie, like the description and rating, impacts the likelihood of passing the Bechdel test.

Introduction:

The Bechdel test is a measure of the inclusion of women in movies based on how they talk. This test examines if a movie has two female characters that have a conversation about anything other than a man at least once throughout the movie. There are several variants of the test, but these are the requirements used by bechdeltest.com, the source of our data.The test was first described in a comic by Alison Bechdel, where a character says "I only go to a movie if it satisfies three basic requirements. One, it has to have at least two women in it... who two, talk to each other about, three, something besides a man." (Dykes to Watch Out For, 1985). Although the test is named after her, Bechdel credits the original idea to her friend, Liz Wallace.

According to a Fivethirtyeight article about the Bechdel test, movies that pass the Bechdel test make more money for each dollar spent compared to movies that don't. However, they are also given the lowest budgets when compared to movies that don't strictly pass the Bechdel test. (Hickey, 2014)

IMDb is a database that contains information about movies, TV shows, video games, actors, directors and producers. IMDb also has data about rating information from viewers and critics as well as a description of the movie. We used this database to add information to our Bechdel data like descriptions and ratings.

Data:

Our bechdeltest.com data is sourced from Fivethirtyeight, and our IMDb data is sourced from Kaggle. It is worth noting that the data provided by Fivethirtyeight (the limiting factor in sourcing our data, as Kaggle had far more movies) represents all of the movies that both bechdeltest.com and the-numbers.com at the time of the creation of the accompanying Fivethirtyeight article (April 2014). As such, it does not represent a specific population of movies besides movies that happened to be in both of these databases. As we are generally using the same data as Fivethirtyeight, we also have the same issue.

We merged our data sets from Fivethirtyeight and Kaggle, and then removed any movies with missing data. In addition, a few movies had descriptions that were cut off in one way or another, so we cut those from the dataset as well. After this, we were left with 1763 movies in our data set, released between the years 1970 and 2013.

After this, we created two custom variables using the IMDb description. "she" represents whether or not a movie has she/her/hers pronouns in its IMDb description, while "he" represents whether or not a movie has he/him/his pronouns in its IMDb description.

```
joined2 <- test %>%
  left_join(IMDbMov, by = "imdb_title_id") %>%
  left_join(IMDbRat, by = "imdb_title_id")

joined <- movies %>%
  left_join(IMDbMov, by = "imdb_title_id") %>%
  left_join(IMDbRat, by = "imdb_title_id")

data <- select(joined, binary, description, males_18age_avg_vote)

data <- data %>%

  mutate(description = ifelse(nchar(description) <= 10 |
                                grepl(",", str_sub(description, -1), fixed = TRUE), NA, description)) %>%


  na.exclude(description)          %>%
  na.exclude(males_18age_avg_vote) %>%

  mutate(ageGroup = cut(males_18age_avg_vote, breaks = 3)) %>%
  mutate(binary = ifelse(binary == "PASS", 1, 0))          %>%

  mutate(she = as.factor(ifelse(grepl(" she " , description, fixed = TRUE) |
                                grepl(" her " , description, fixed = TRUE) |
                                grepl(" hers ", description, fixed = TRUE) |
                                grepl("She "  , description, fixed = TRUE) |
                                grepl("Her "  , description, fixed = TRUE) |
                                grepl(" she." , description, fixed = TRUE) |
                                grepl(" her." , description, fixed = TRUE) |
                                grepl(" hers.", description, fixed = TRUE) |
                                grepl(" she," , description, fixed = TRUE) |
                                grepl(" her," , description, fixed = TRUE) |
                                grepl(" hers,", description, fixed = TRUE) , 1, 0))) %>%

  mutate(he = as.factor(ifelse( grepl(" he "  , description, fixed = TRUE) |
                                grepl(" him " , description, fixed = TRUE) |
                                grepl(" his " , description, fixed = TRUE) |
                                grepl("He "   , description, fixed = TRUE) |
                                grepl("His "  , description, fixed = TRUE) |
                                grepl(" he."  , description, fixed = TRUE) |
                                grepl(" his." , description, fixed = TRUE) |
                                grepl(" him." , description, fixed = TRUE) |
                                grepl(" he,"  , description, fixed = TRUE) |
                                grepl(" him," , description, fixed = TRUE) |
                                grepl(" his," , description, fixed = TRUE) , 1, 0)))
```

Model:

```
model <- glm(binary ~ males_18age_avg_vote + she + he, data = data, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = binary ~ males_18age_avg_vote + she + he, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3035  -0.9950  -0.6891   1.1825   1.9126
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.43299    0.38817   3.692 0.000223 ***
## males_18age_avg_vote -0.24081    0.05647  -4.264 2.00e-05 ***
## she1                   1.82121    0.15796  11.530  < 2e-16 ***
## he1                   -0.89548    0.11112  -8.059 7.71e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2422.9  on 1762  degrees of freedom
## Residual deviance: 2112.6  on 1759  degrees of freedom
## AIC: 2120.6
##
## Number of Fisher Scoring iterations: 4
```

```
bechdel_model <- augment(model, data = data)
bechdel_model <- bechdel_model %>%
  mutate(odds = exp(.fitted),
         probability = odds / (1 + odds))

bechdel_model <- bechdel_model %>%
  mutate(predictPass = if_else(probability >= 0.5, 1, 0))

exp(coef(model))
```

```
##          (Intercept) males_18age_avg_vote                 she1
##            4.1912139            0.7859874            6.1793529
##                  he1
##            0.4084103
```

The explanatory variables we used for our model were "she", "he", and "males_18age_avg_vote"

Males_18age_avg_vote is the average rating of a male between 18-29 years old for a given movie. There were also average ratings for either sex (male or female) split into their respective age groups, all ages split into either sex, and the mean and median ratings overall. The age groups used were 0-17, 18-29, 30-44, and 45+. We decided to use males_18age_avg_vote because it was the most statistically significant and had

the highest effect on our model and we could only use one variable from the set of ratings by demographic as they were generally highly correlated with each other and so we would have problems with multicollinearity.

All of variables had statistically significant P-values (<0.05)

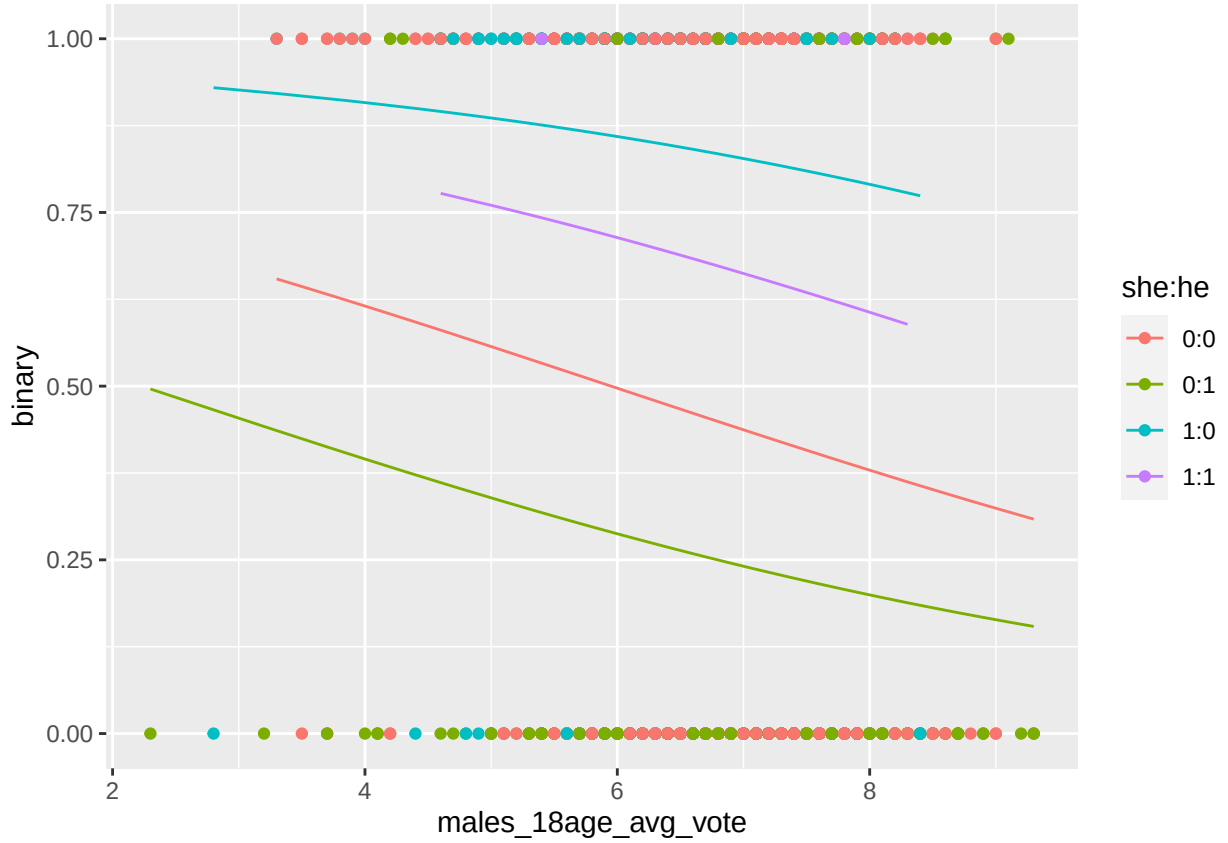
Looking at these coefficients, we can come to the following conclusions:

A 1-point increase in the average rating by males age 18 to 29 is associated with multiplying the odds of passing the test by a factor of 0.785, holding all else constant, in other words the odds decrease.

Compared to movies where she/her/hers pronouns were not present, the odds of a movie with she/her/hers pronouns passing the Bechdel test were multiplied by a factor of 6.179, holding all else constant. In other words the odds increase.

Compared to movies where he/him/his pronouns were not present, the odds of a movie with he/him/his pronouns passing the Bechdel test were multiplied by a factor of 0.408, holding all else constant. In other words the odds decrease.

```
ggplot(bechdel_model, aes(x = males_18age_avg_vote, color = she : he)) +
  geom_point(aes(y = binary)) +
  geom_line(aes(y = probability))
```



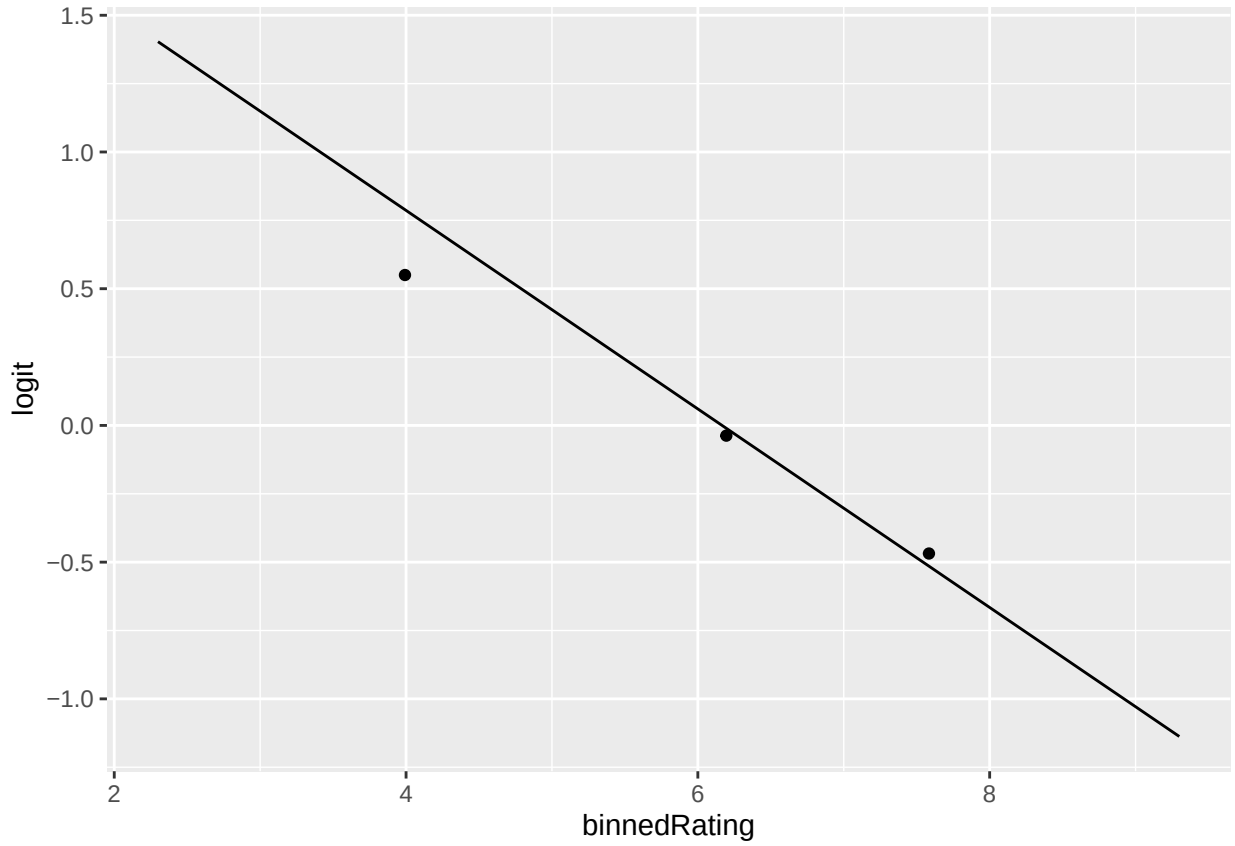
Tests:

```
malesm1 <- glm(binary ~ males_18age_avg_vote, data = data, family = binomial)
```

```
malestestm1 <- augment(malesm1, data = data)
malestestm1 <- malestestm1 %>%
  mutate(odds = exp(.fitted),
         probability = odds / (1 + odds))

bechdel_binned <- data %>%
  group_by(ageGroup) %>%
  summarize(binnedRating = mean(males_18age_avg_vote), binnedPass = mean(binary)) %>%
  mutate(logit = log(binnedPass/(1-binnedPass)))

ggplot(bechdel_binned) +
  geom_point(aes(x = binnedRating, y = logit)) +
  geom_line(data = malestestm1, aes(x = males_18age_avg_vote, y = .fitted))
```



We used an empirical logit plot to check if our model is sufficiently linear. We used 3 bins, and from the plot produced we saw that the line was sufficiently linear and the points were well fitted to the line. Based on this, we determined that the linearity condition for our model is upheld.

```
vif(model)
```

```
## males_18age_avg_vote                  she                   he
##             1.008121             1.006891             1.004486
```

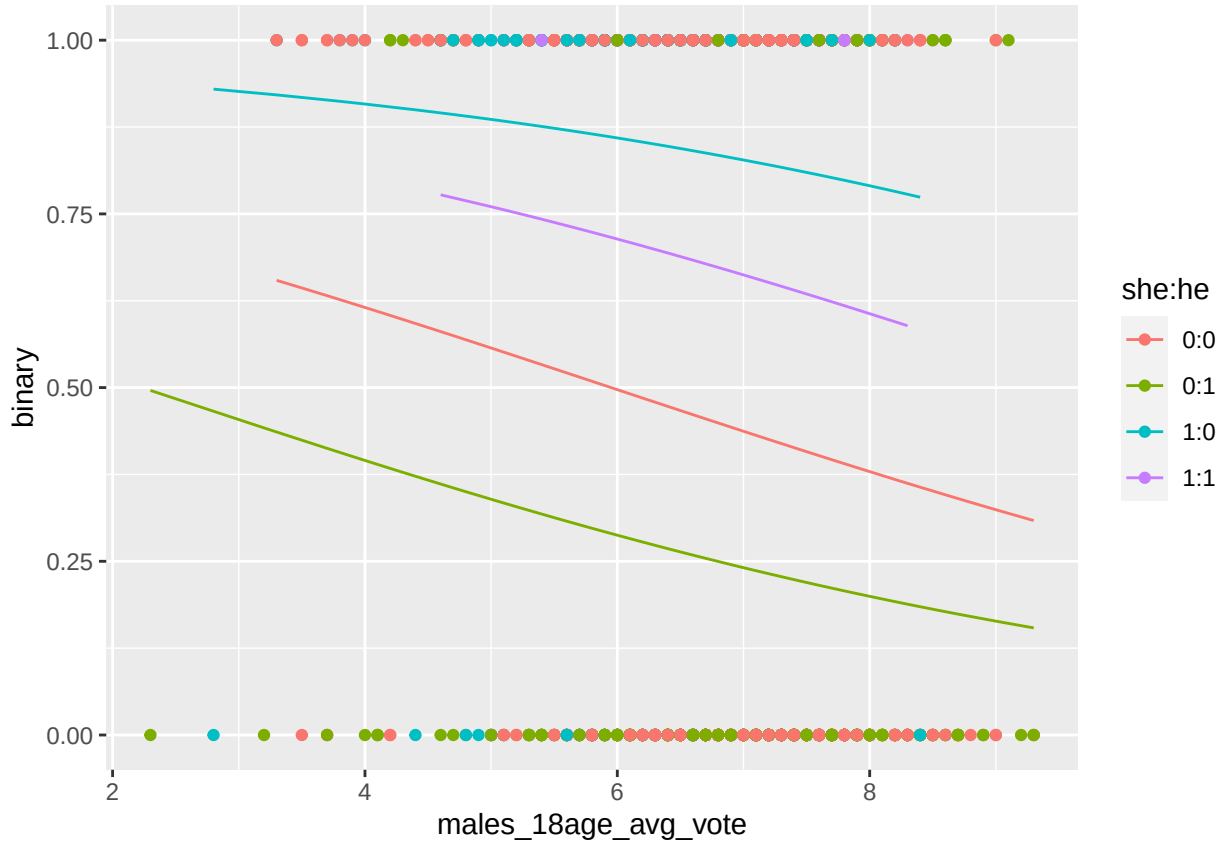
```
Hmisc::rcorrcens(binary ~ fitted.values(model), data = data)
```

```
##
## Somers' Rank Correlation for Censored Data    Response variable:binary
##
##                           C   Dxy  aDxy    SD     Z P    n
## fitted.values(model) 0.728 0.456 0.456 0.024 18.68 0 1763
```

Our VIF test shows that we do not have issues with multicollinearity between our explanatory variables- all of the values are less than 5.

In order to test the quality of the fit of our model, we found the C-statistic. For our model, this was 0.72, and had a P-value of ~0. As our P-value was significant, we decided that our model has a sufficiently high quality of fit.

```
ggplot(bechdel_model, aes(x = males_18age_avg_vote, color = she : he)) +
  geom_point(aes(y = binary)) +
  geom_line(aes(y = probability))
```



We graphed the probabilities of a movie passing the Bechdel test based on the average rating from a 18-29 year old male and the categorical variables she and he, as described above. This results in a negative linear trend. The she:he variables shown on the side are represented as either 0:0, 0:1, 1:0 or 1:1. A "1" represents that pronoun as being present in the description so the highest probability of passing the Bechdel test was she pronouns were present and he pronouns weren't. This is followed by both types of pronouns being present in the description. Third is when neither pronouns were mentioned. And lastly, the lowest chance of passing the Bechdel test is if traditional female and male pronouns aren't present. From this graph, you can

also see the negative correlation between the rating a movie received from males ages 18-29 and whether or not a movie passed the Bechdel test, regardless of the contents of the description.

```
bechdel_model %>%
  group_by(binary, predictPass) %>%
  summarize(n())
```

```
## `summarise()` has grouped output by 'binary'. You can override using the `.groups` argument.
```

```
## # A tibble: 4 x 3
## # Groups:   binary [2]
##   binary predictPass `n()`
##    <dbl>       <dbl> <int>
## 1      0           0   861
## 2      0           1   117
## 3      1           0   442
## 4      1           1   343
```

```
avg <- mean(data$binary)
1-avg
```

```
## [1] 0.5547362
```

```
(861+343)/(862+343+117+442)
```

```
## [1] 0.6825397
```

Based on the numbers shown above, we are able to accurately predict if a movie passes the Bechdel test 68.27% of the time. If we used the mean to predict whether or not a movie will pass the Bechdel test, we would predict that every movie would fail the Bechdel test, making us right only 55.47% of the time.

Conclusion:

After combining the data sets and seeing the impact several variables had on the chance of passing the Bechdel test, we came to the conclusion that we can predict whether or not a movie passes the Bechdel test based on the average ratings from 18-29 year old men and the IMDb description with higher accuracy than using the mean. However, as our data does not represent a specific population of movies, we do not have a population to extrapolate our results to.

Additionally, it should be mentioned, "the Bechdel test is not necessarily the best benchmark to measure female representation in movies. It doesn't take into consideration how well written a character is, neither does it measure meaningful depth of character." (Selvaraj, 2020) The Bechdel test is a rather low bar to measure female representation but is definitive to measure and is interesting to talk about because of the surprising number of movies that don't pass.

Works Cited

Fivethirtyeight. "Data/Movies.csv at Master · Fivethirtyeight/Data." GitHub, github.com/fivethirtyeight/data/blob/master/bechdel/movies.csv.

Selvaraj, Natassha. "The Bechdel Test: Analyzing Gender Disparity in Hollywood." Medium, Towards Data Science, 5 June 2020, towardsdatascience.com/the-bechdel-test-analyzing-gender-disparity-in-hollywood-263cd4bcd9d.

WaltHickey. "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women." FiveThirtyEight, FiveThirtyEight, 1 Apr. 2014, fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/.

stefanoleone992. "IMDb extensive dataset." Kaggle, https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset (link no longer functional)