Analyzing the Bechdel Test: Budget Trends and Revenue Outcomes in Cinema

December 15th 2024

Alex Ackeman, Cindy Gao, Kayla Haeussler, Javidan Karimli

1. Abstract

Understanding the factors influencing gender representation in films is essential for addressing disparities in media. This study examines the relationship between movie budgets and their likelihood of passing the Bechdel Test, as well as differences in international box office revenue between movies that pass or fail the test. The modulating effect of decade and genre is also explored for its implications on media representation and industry practices. The dataset analyzed is derived from BechdelTest.com, captured from the 'fivethirtyeight' repository on GitHub. Movie genres were applied to the dataset using the IMDb library in Python. Logistic regression was applied to model Bechdel Test outcomes, and linear regression was used to model international box office revenue. A priori variable selection and exploratory data analysis were conducted for variable inclusion. The interaction between inflation-adjusted budget and decade was used to explore changes over time. Genre was included as a covariate when modeling international revenue.

This study provides insights into the economic and cultural dynamics shaping gender representation in the film industry and serves as a reference for stakeholders seeking to promote inclusivity in media production.

2. Introduction

Gender representation in media is a critical issue due to its influence on societal norms and perceptions. The Bechdel Test, a measure of whether movies include at least two named women characters who talk to each other about something other than a man, highlights persistent disparities in representation [1]. Despite growing awareness, many films continue to fail this basic measure of inclusivity, raising questions about the social and economic factors that contribute to this outcome.

The film industry is a major global enterprise, with Hollywood alone generating over $100 billion in revenue annually [2]. Budget allocations, genre choices, and audience preferences significantly influence the production and success of movies. Previous studies suggest that higher-budget movies often prioritize traditional narratives that may not meet diversity benchmarks like the Bechdel Test [3]. Conversely, films that pass the test have demonstrated competitive performance at the box office, particularly in international markets, suggesting that inclusivity can align with financial success [4].

Economic and cultural factors, such as inflation-adjusted budgets, genre conventions, and evolving societal norms across decades, may influence gender representation in films. By analyzing a dataset from BechdelTest.com, which includes information on movies' budgets, box office performance, genres (applied using the IMDb library in Python), and Bechdel Test outcomes, this study seeks to address the following questions:

1. What is the relationship between a movie's budget and its likelihood of passing the Bechdel Test? Does this relationship vary across decades?
2. How does passing the Bechdel Test impact a movie's international box office revenue, and does this relationship depend on the movie's genre?

By examining these questions, the study aims to provide insights into the economic and cultural dynamics shaping gender representation in the film industry and contribute to ongoing discussions on inclusivity in media.

3. Methods

3.1 Data and Preprocessing

Our dataset was obtained from the fivethirtyeight article *The Dollar-And-Cents Case Against Hollywood's Exclusion of Women,* which synthesized Bechdel test result data, from BechdelTest.com, and movie financial metrics, from TheNumbers.com. The dataset contains 1,794 rows and 15 columns, with each row representing a movie, whether or not it passed the Bechdel test, as well as its budget, domestic and international revenue, both in release year dollars and adjusted for inflation to 2013 dollars.

In addition to the information provided in the original dataset, our team thought it would be interesting to explore movie genre as a variable as well. The original dataset includes a column labeled imdb, containing unique identifiers corresponding to movie details available in the IMDb database. To extract and classify movie genres, the PyMovieDb library was utilized. Genres were categorized into seven distinct groups: Comedy, Romance, Action, Horror, Drama, Family, and Other. If a movie's genre did not align with any of the first six categories, it was automatically assigned to the "Other" category. Given that many movies belong to multiple genres, the classification process prioritized the first match among the predefined categories.

To begin our data processing, we removed the columns for budget, domestic revenue and international revenue which had these amounts in US dollar at time of release, and retained the columns which had these values in 2013 USD values. Our team felt it best to use only the 2013 inflation adjusted columns for each of these values in our analysis to ensure consistent and equitable comparison of financial metrics. Our initial exploration of the data also revealed missing values in the domestic and international gross revenue (2013) columns, as well as, most notably, the decade code column, which assigns a single digit corresponding to the movie's release decade. There were 179 rows of our data which were missing a decade code value, which we discovered were all films released in the 1970s and 1980s. We edited the data to fill in a '4' and '5' for movies released in the 1980s and 1970s, respectively. As far as data missing in the domestic and gross revenue columns, we could not pin point the exact cause of the data not being available. There were 18 films missing domestic gross revenue (2013) data, with 11 of those also missing international gross revenue data. We thought this may be due to international movies being included in the data set, but this was not the case. Due to this only being a small number of rows in our dataset and the reason for them missing being unclear, we felt it best to remove these rows missing domestic and international revenue from our dataset.

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(tidymodels)
library(ISLR2)
library(car)
```

```
# load the raw data and format all the columns
movie_data <- read_csv(
  "https://raw.githubusercontent.com/ackerman-alex/IDS_702_Final_Project/refs/heads/main/Mov
  col_types = cols(
    imdb = col_character(),           # IMDb identifier
    title = col_character(),          # Movie title
    year = col_double(),              # Year of release
    test = col_character(),           # Original Bechdel Test result
    clean_test = col_character(),     # Cleaned-up test result
    binary = col_character(),         # PASS/FAIL outcome
    budget = col_double(),            # Original budget
    domgross = col_double(),          # Original domestic gross
    intgross = col_double(),          # Original international gross
    code = col_character(),           # Classification code
    `budget_2013$` = col_double(),    # Inflation-adjusted budget
    `domgross_2013$` = col_double(),  # Inflation-adjusted domestic gross
    `intgross_2013$` = col_double(),  # Inflation-adjusted international gross
    `period code` = col_double(),     # Period code
```

```
    `decade code` = col_double(),     # Decade code
    genre = col_character()           # Movie genre
  )
)

# rename the columns to get rid of the space in the names.
movie_data <- movie_data %>%
  rename(
    budget_2013 = `budget_2013$`,
    domgross_2013 = `domgross_2013$`,
    intgross_2013 = `intgross_2013$` ,
    period_code = `period code`,
    decade_code = `decade code`
  )

# removing period code column because idk what it is and also it has a lot of missing values
movie_data$period_code <- NULL

head(movie_data)
```

```
# A tibble: 6 x 15
   year imdb      title   test  clean_test binary budget domgross intgross code
  <dbl> <chr>     <chr>   <chr> <chr>      <chr>   <dbl>    <dbl>    <dbl> <chr>
1  2013 tt1711425 21 &am~ nota~ notalk     FAIL   1.3 e7 25682380   4.22e7 2013~
2  2012 tt1343727 Dredd ~ ok-d~ ok         PASS   4.50e7 13414714   4.09e7 2012~
3  2013 tt2024544 12 Yea~ nota~ notalk     FAIL   2   e7 53107035   1.59e8 2013~
4  2013 tt1272878 2 Guns  nota~ notalk     FAIL   6.1 e7 75612460   1.32e8 2013~
5  2013 tt0453562 42      men   men        FAIL   4   e7 95020213   9.50e7 2013~
6  2013 tt1335975 47 Ron~ men   men        FAIL   2.25e8 38362475   1.46e8 2013~
# i 5 more variables: budget_2013 <dbl>, domgross_2013 <dbl>,
#   intgross_2013 <dbl>, decade_code <dbl>, genre <chr>
```

```
# check for missing values
colSums(is.na(movie_data))[colSums(is.na(movie_data)) > 0]
```

```
     domgross      intgross domgross_2013 intgross_2013   decade_code
           17            11            18            11           179
```

```
movie_data <- movie_data %>%
  mutate(decade_code = as.numeric(decade_code))
```

```r
# Fill with the missing value:
# for movies from 1980-1989 (included), the decade code is 4.
# for movies from 1970-1979 (included), the decade code is 5.

movie_data <- movie_data %>%
  mutate(
    decade_code = case_when(
      is.na(decade_code) & year >= 1980 & year <= 1989 ~ 4.0,
      is.na(decade_code) & year >= 1970 & year <= 1979 ~ 5.0,
      TRUE ~ decade_code
    )
  )

movie_data <- na.omit(movie_data)
nrow(movie_data)
```

```
[1] 1776
```

```r
movie_data <- movie_data %>%
  mutate(
    binary = factor(binary, levels = c("FAIL", "PASS")),
    decade_code = factor(decade_code)
  )
```

3.2 Variable Selection

3.3 Model Fitting and Evaluation

4. Results

4.1 Overview of Included Data

After initial data preprocessing, our dataset contains 1,776 rows. Of these 1,776 films, 794 of them passed the Bechdel test, while 982 of the films passed the test.

Table 1: Summary Statistics of Movie Financial Characteristics in 2013 USD

	Budget 2013	Domestic Gross Revenue 2013	International Gross Revenue 2013
Median	36995786	55993640	96239640

	Budget 2013	Domestic Gross Revenue 2013	International Gross Revenue 2013
1st Quartile	16068918	20546594	33232604
3rd Quartile	78337905	121678352	241478970
Mean	55464608	95174784	197837985

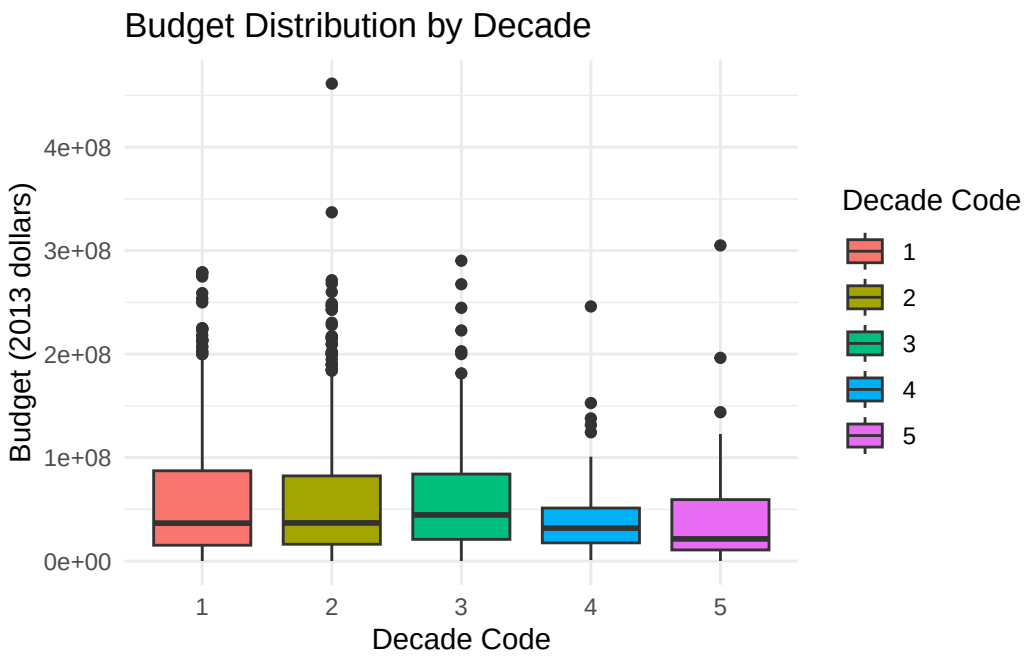
4.2 Research Question 1: Relationship Between the Bechdel Test and Budget

The results of the logistic regression model are shown in table 2 below

Table 2: Logistic Regression Model Summary

Variable	Estimate	Std Error	t-value	p-value

```
# box plot for Budget Distribution by Decade
ggplot(movie_data, aes(x = as.factor(decade_code), y = budget_2013, fill = as.factor(decade_
  geom_boxplot() +
  labs(title = "Budget Distribution by Decade",
       x = "Decade Code",
       y = "Budget (2013 dollars)",
       fill = "Decade Code") +
  theme_minimal()
```



Interpretation: The overall budget levels have remained relatively stable across decades when adjusted for inflation. Since movie budgets do not differ significantly across decades, the interaction term between budget_2013 and decade_code in the regression model may help explain decade-specific effects on passing the Bechdel Test.

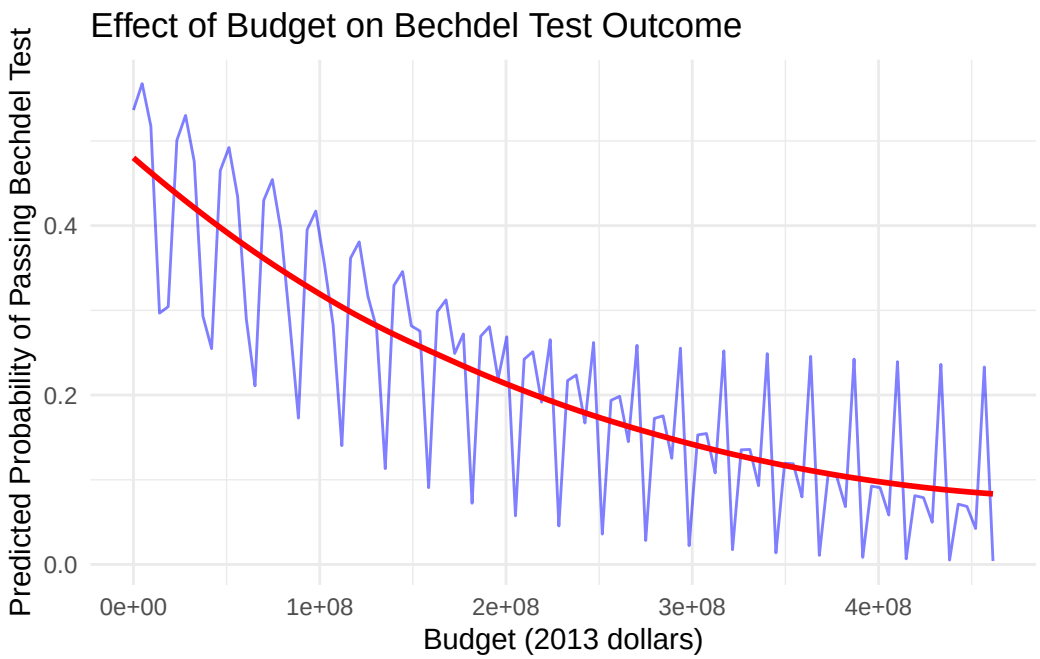
```
# Construct the logistic regression model
glm_model <- glm(binary ~ budget_2013 * decade_code, data = movie_data, family = "binomial")
summary(glm_model)
```

```
Call:
glm(formula = binary ~ budget_2013 * decade_code, family = "binomial",
    data = movie_data)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.461e-01  1.365e-01   1.070 0.284543
budget_2013               -6.132e-09  1.705e-09  -3.596 0.000323 ***
decade_code2               1.568e-01  1.704e-01   0.920 0.357647
decade_code3              -8.124e-03  2.218e-01  -0.037 0.970787
decade_code4              -9.987e-01  3.358e-01  -2.974 0.002939 **
decade_code5              -7.755e-01  4.404e-01  -1.761 0.078281 .
budget_2013:decade_code2 -3.768e-10  2.197e-09  -0.172 0.863803
```

```
budget_2013:decade_code3 -1.069e-09  3.065e-09  -0.349 0.727371
budget_2013:decade_code4  5.391e-09  6.095e-09   0.884 0.376440
budget_2013:decade_code5 -4.444e-09  8.883e-09  -0.500 0.616850
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2442.1  on 1775  degrees of freedom
Residual deviance: 2367.9  on 1766  degrees of freedom
AIC: 2387.9

Number of Fisher Scoring iterations: 4
```

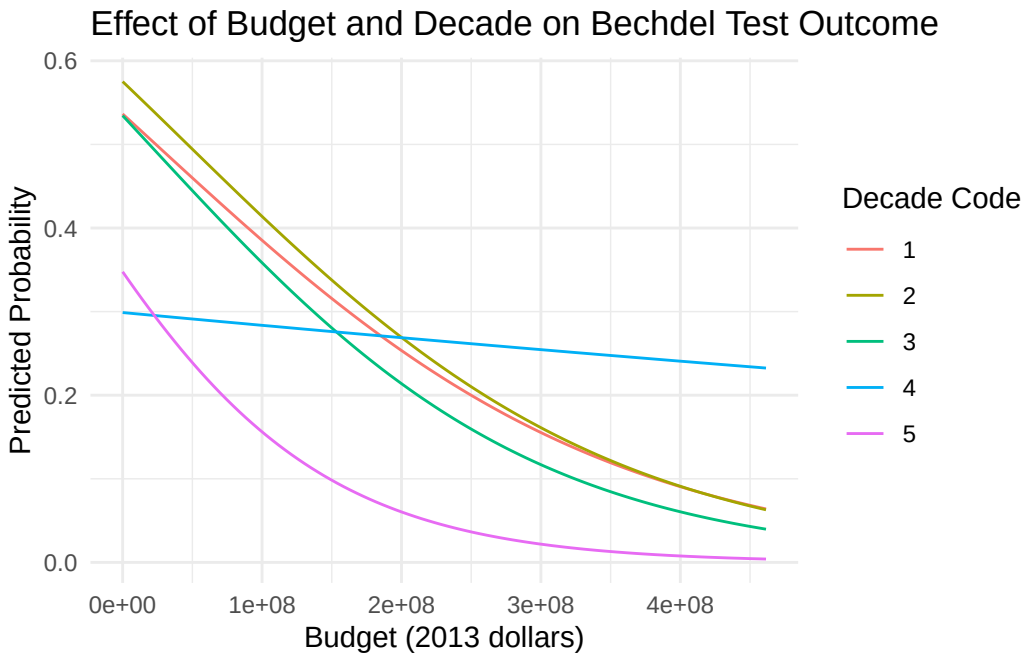
```r
new_data <- data.frame(
  budget_2013 = seq(min(movie_data$budget_2013, na.rm = TRUE),
                    max(movie_data$budget_2013, na.rm = TRUE),
                    length.out = 100),
  decade_code = levels(movie_data$decade_code)
)


new_data$predicted_prob <- predict(glm_model, newdata = new_data, type = "response")


ggplot(new_data, aes(x = budget_2013, y = predicted_prob)) +
  geom_line(color = "blue", alpha = 0.5) +
  geom_smooth(method = "loess", color = "red", se = FALSE) +
  labs(title = "Effect of Budget on Bechdel Test Outcome",
       x = "Budget (2013 dollars)",
       y = "Predicted Probability of Passing Bechdel Test") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Thoughts: **Movies with higher budgets have a lower predictive probability of passing the Bechdel test.**

```
new_data <- expand.grid(
  budget_2013 = seq(min(movie_data$budget_2013, na.rm = TRUE),
                    max(movie_data$budget_2013, na.rm = TRUE),
                    length.out = 100),
  decade_code = unique(movie_data$decade_code)
)


new_data$predicted_prob <- predict(glm_model, newdata = new_data, type = "response")


ggplot(new_data, aes(x = budget_2013, y = predicted_prob, color = factor(decade_code))) +
  geom_line() +
  labs(title = "Effect of Budget and Decade on Bechdel Test Outcome",
       x = "Budget (2013 dollars)",
       y = "Predicted Probability",
       color = "Decade Code") +
  theme_minimal()
```

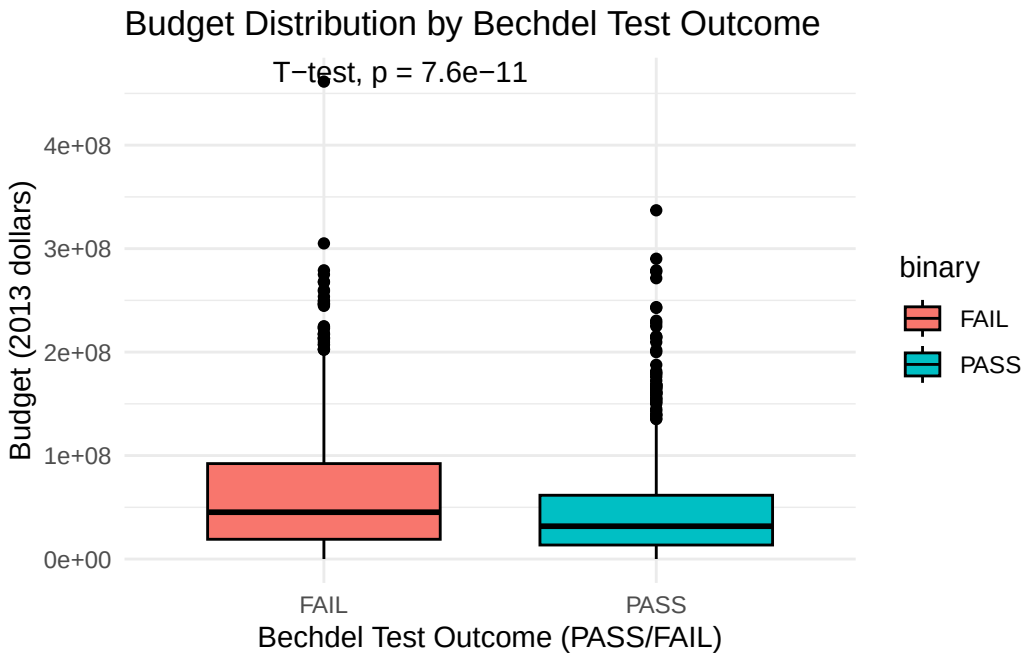


Interpretation:

For Decade Code 1, 2, 3, and 5: The trends are similar—higher budgets are associated with a lower probability of passing the Bechdel Test. This might indicate that high-budget films in these decades are more focused on genres like action or sci-fi, which are less likely to pass the test.

For Decade Code 4: The curve is nearly flat, meaning budget has little impact on passing the Bechdel Test. This could suggest that high-budget films in this decade are less influenced by genre differences in gender representation.

```
# box plot for Budget Distribution by Bechdel Test Outcome
library(ggpubr)

ggboxplot(movie_data, x = "binary", y = "budget_2013", fill = "binary") +
  stat_compare_means(method = "t.test") +
  labs(title = "Budget Distribution by Bechdel Test Outcome",
       x = "Bechdel Test Outcome (PASS/FAIL)",
       y = "Budget (2013 dollars)") +
  theme_minimal()
```



Interpretation:

- **p-value**: The t-test p-value is 7.6e-11, which is much smaller than 0.05, which indicates that there is a significant difference in budget distribution between the two groups, suggesting that budget may be an important factor affecting Bechdel Test outcomes.

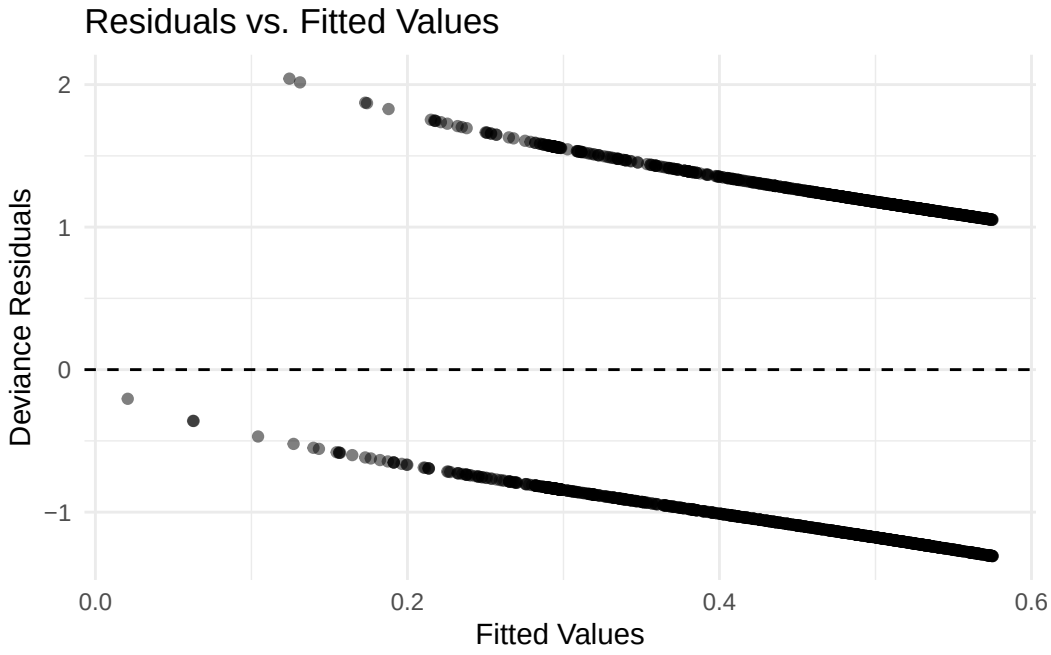
Overall:

- High-budget movies are more likely to fail the test, and low-budget movies are more likely to pass the test.
- Some possible explanation would be related to **Movie Genre**: High-budget movies, such as action or sci-fi, often have weaker gender representation and are less likely to pass the test. In contrast, low-budget movies, like dramas or independent films, tend to focus more on gender equality and are more likely to pass.

Some ways to refine the current glm:

```
residuals <- residuals(glm_model, type = "deviance")

# Residuals vs. Fitted Values
ggplot(data = data.frame(fitted = fitted(glm_model), residuals = residuals),
       aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed") +
```

```
  labs(title = "Residuals vs. Fitted Values",
       x = "Fitted Values",
       y = "Deviance Residuals") +
  theme_minimal()
```

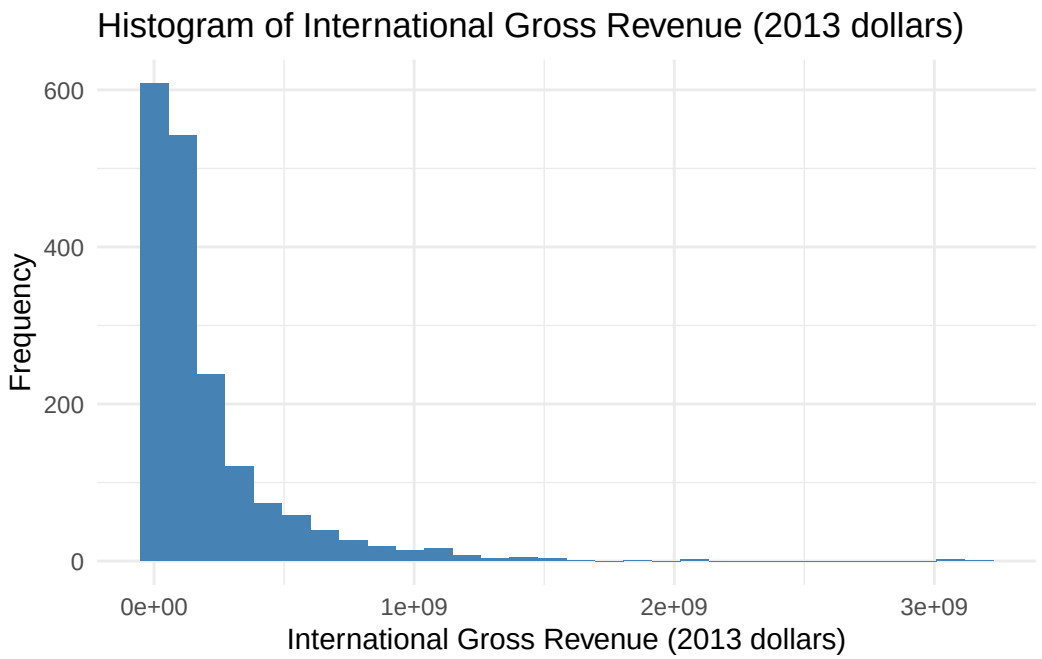


Thoughts: I think the model here needs to be refined because there is a certain trend, ideally we want the residuals should be randomly distributed around 0. Therefore, we probably need to add more predictors to form a more detailed model to compare with the current one.

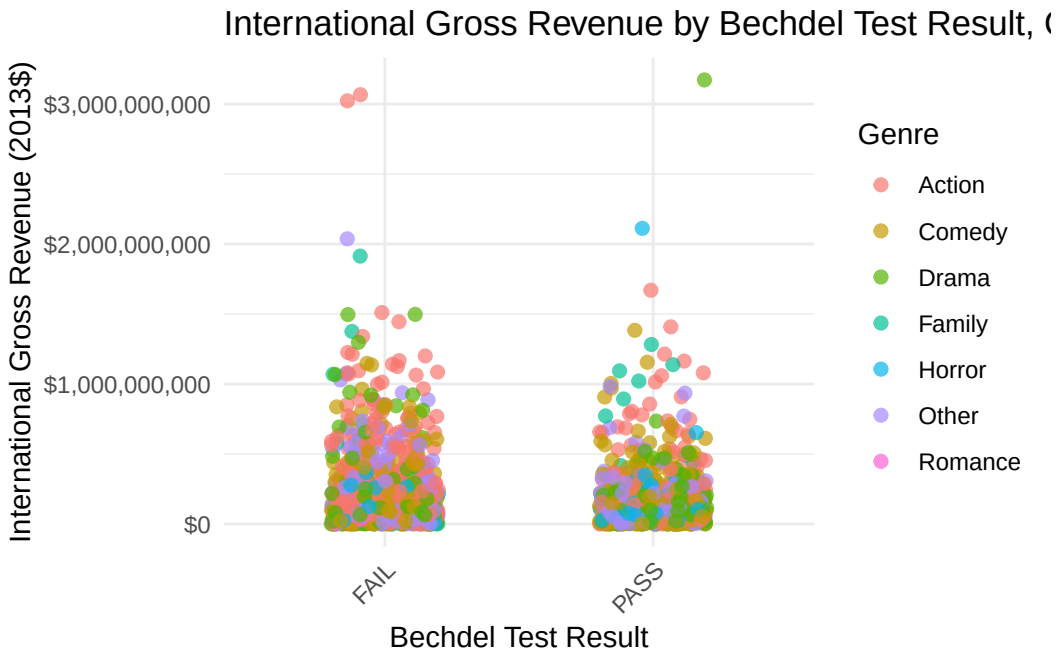
4.3 Research Question 2: Bechdel Test, Genre, and the International Box Office

	Action	Comedy	Drama	Family	Horror	Romance	Other
Number of Movies in Dataset	483	498	442	23	100	2	46

```
# OUTCOME VAR: Histogram for international gross revenue
ggplot(movie_data, aes(x = `intgross_2013`)) +
  geom_histogram(fill = "steelblue", bins = 30) +
  labs(x = "International Gross Revenue (2013 dollars)", y = "Frequency") +
  ggtitle("Histogram of International Gross Revenue (2013 dollars)") +
  theme_minimal()
```




```r
# Convert the intgross_2013$ column to numeric after removing any commas
movie_data$intgross_2013_numeric <- as.numeric(gsub(",", "", movie_data$`intgross_2013`))

# Scatter plot with genre coloring and Bechdel Test result on x-axis
ggplot(movie_data, aes(x = binary, y = intgross_2013_numeric, color = genre)) +
  geom_jitter(position = position_jitter(width = 0.2, height = 0), alpha = 0.7, size = 2) +
  scale_y_continuous(labels = scales::dollar) +
  labs(
    title = "International Gross Revenue by Bechdel Test Result, Colored by Genre",
    x = "Bechdel Test Result",
    y = "International Gross Revenue (2013$)",
    color = "Genre"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
ResearchQ2Model <- lm(intgross_2013 ~  binary  + budget_2013 +domgross_2013+ decade_code + g

summary(ResearchQ2Model)
```

```
Call:
lm(formula = intgross_2013 ~ binary + budget_2013 + domgross_2013 +
    decade_code + genre, data = movie_data)

Residuals:
       Min         1Q     Median         3Q        Max
-657169381  -39253986    1628529   30333779 1039355894

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.698e+07  6.673e+06  -4.042 5.52e-05 ***
binaryPASS     8.369e+06  4.321e+06   1.937 0.052962 .
budget_2013    9.505e-01  4.747e-02  20.022  < 2e-16 ***
domgross_2013  1.972e+00  2.023e-02  97.440  < 2e-16 ***
decade_code2  -2.012e+07  5.174e+06  -3.888 0.000105 ***
decade_code3  -2.219e+07  6.380e+06  -3.479 0.000516 ***
decade_code4  -7.788e+07  9.241e+06  -8.428  < 2e-16 ***
```

```
decade_code5  -1.563e+08  1.363e+07 -11.465  < 2e-16 ***
genreComedy   -2.991e+05  6.016e+06  -0.050 0.960352
genreDrama     1.279e+07  6.230e+06   2.053 0.040267 *
genreFamily    8.590e+07  1.875e+07   4.581 4.95e-06 ***
genreHorror    1.662e+07  1.015e+07   1.637 0.101708
genreOther    -2.830e+06  7.028e+06  -0.403 0.687241
genreRomance  -6.969e+06  6.175e+07  -0.113 0.910159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 86970000 on 1762 degrees of freedom
Multiple R-squared:  0.9068,    Adjusted R-squared:  0.9061
F-statistic:  1319 on 13 and 1762 DF,  p-value: < 2.2e-16
```

```
confint(ResearchQ2Model)
```

```
                    2.5 %        97.5 %
(Intercept)   -4.006588e+07 -1.388847e+07
binaryPASS    -1.070858e+05  1.684450e+07
budget_2013    8.574099e-01  1.043635e+00
domgross_2013  1.931985e+00  2.011358e+00
decade_code2  -3.026545e+07 -9.969000e+06
decade_code3  -3.470588e+07 -9.680019e+06
decade_code4  -9.600649e+07 -5.975875e+07
decade_code5  -1.830023e+08 -1.295345e+08
genreComedy   -1.209934e+07  1.150107e+07
genreDrama     5.682320e+05  2.500772e+07
genreFamily    4.912477e+07  1.226792e+08
genreHorror   -3.287398e+06  3.653369e+07
genreOther    -1.661465e+07  1.095453e+07
genreRomance  -1.280843e+08  1.141463e+08
```

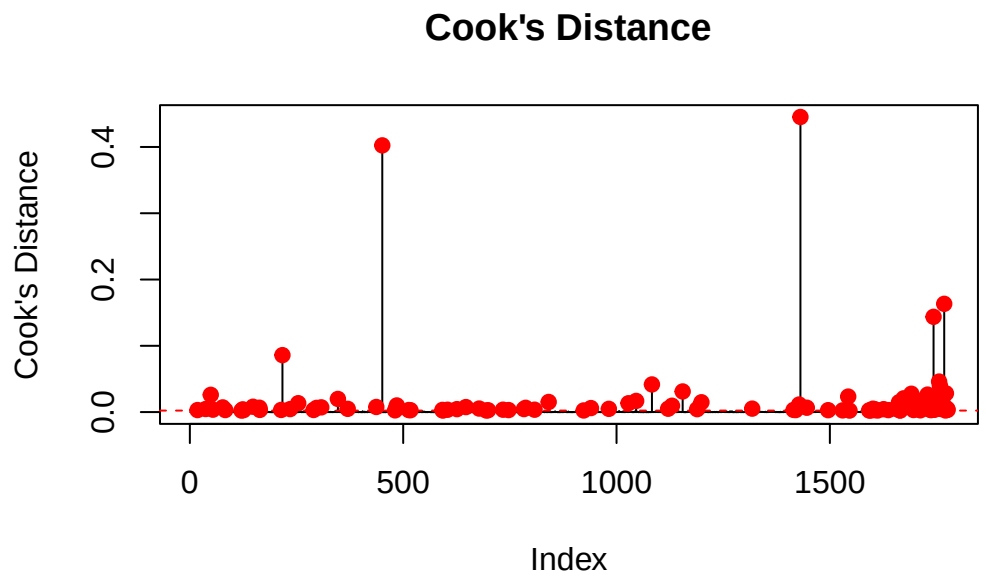
```
vif(ResearchQ2Model)
```

```
                 GVIF Df GVIF^(1/(2*Df))
binary       1.083932  1        1.041120
budget_2013  1.601050  1        1.265326
domgross_2013 1.524540  1        1.234722
decade_code  1.274055  4        1.030739
genre        1.338877  6        1.024617
```

Cook's Distance was evaluated to detect any extreme values in the data. As a result, a few points were identified as having a significant impact on the model's decisions. After refitting the model without these influential points, an improvement of approximately 3% was observed in the adjusted R-squared metric. Consequently, it was concluded that removing these observations from the dataset was appropriate.

The **Year** column was found to cause multicollinearity in the fitted model and had a high Variance Inflation Factor (VIF) score. Upon careful investigation of the data, it was discovered that the **Decade** code conveys the same information as the **Year** column but provides a more generalized approach. Therefore, the removal of the **Year** column was decided to improve model performance and reduce multicollinearity.

```
cooks_distances <- cooks.distance(ResearchQ2Model)
n <- nrow(movie_data)
threshold <- 4 / n
influential_points <- which(cooks_distances > threshold)

plot(cooks_distances, type = "h", main = "Cook's Distance",
     ylab = "Cook's Distance", xlab = "Index")

abline(h = threshold, col = "red", lty = 2)

influential_indices <- which(cooks_distances > threshold)
points(influential_indices, cooks_distances[influential_indices], col = "red", pch = 19)
```



```r
movie_data_no_influential <- movie_data[-influential_points, ]
ResearchQ2ModelWithoutExtreme <- lm(intgross_2013 ~ + binary + genre  + budget_2013 +domgross

summary(ResearchQ2ModelWithoutExtreme)
```

```
Call:
lm(formula = intgross_2013 ~ +binary + genre + budget_2013 +
    domgross_2013 + decade_code, data = movie_data_no_influential)

Residuals:
       Min         1Q     Median         3Q        Max
-187785936  -29053463    1302788   24897316  222099269

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.181e+07  4.148e+06  -2.846 0.004485 **
binaryPASS     1.087e+06  2.683e+06   0.405 0.685325
genreComedy   -7.988e+05  3.721e+06  -0.215 0.830069
genreDrama     7.086e+06  3.858e+06   1.837 0.066400 .
genreFamily    7.381e+07  1.996e+07   3.698 0.000224 ***
genreHorror    4.407e+06  6.284e+06   0.701 0.483162
genreOther    -4.196e+06  4.451e+06  -0.943 0.346000
budget_2013    6.790e-01  3.385e-02  20.060  < 2e-16 ***
domgross_2013  1.961e+00  1.699e-02 115.426  < 2e-16 ***
decade_code2  -1.802e+07  3.182e+06  -5.661 1.77e-08 ***
decade_code3  -2.477e+07  3.920e+06  -6.320 3.36e-10 ***
decade_code4  -6.439e+07  5.996e+06 -10.739  < 2e-16 ***
decade_code5  -1.278e+08  1.168e+07 -10.941  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52220000 on 1645 degrees of freedom
Multiple R-squared:  0.9372,    Adjusted R-squared:  0.9367
F-statistic:  2044 on 12 and 1645 DF,  p-value: < 2.2e-16
```

5. Conclusion

6. References

1. Bechdel, A. (1985). Dykes to Watch Out For. *First published as a comic strip in The Essential Dykes to Watch Out For.*
2. Motion Picture Association. (2023). *THEME Report: A Comprehensive Analysis of the Global Film Industry.*
3. Smith, S. L., Choueiti, M., & Pieper, K. (2022). *Inequality in 1,300 Popular Films: Examining Gender, Race, & Ethnicity.* USC Annenberg Inclusion Initiative.
4. Lauzen, M. M. (2021). *The Celluloid Ceiling: Behind-the-Scenes Employment of Women on the Top 250 Films of 2020.* Center for the Study of Women in Television & Film, San Diego State University.