The Impact of the MeToo Movement on IMDb Ratings

Tandem J. Young & Shreyash Gupta

Sam M. Walton College of Business, University of Arkansas

ECON 5783: Applied Microeconometrics

Dr. Arya Gaduh

December 13, 2023

Abstract

In our paper we seek to shed light on the behavior of movie ratings following exogenous events in the film industry. It is through using the comprehensive IMDb database and Difference-in-Difference techniques that we seek to find causality through this behavior. We found impact of women-led films receiving lower than average ratings, with excellent parallel trends and evidence of another shock in 2020.

Table of Contents

Introduction

Accumulating over $77 billion in revenue during 2022, the global film industry is one of the biggest drivers of consumer entertainment. An industry this large and lucrative is not without its faults. In 2017, *The New York Times* published an investigative report into producer Harvey Weinstein. This report outlined over two decades of predatory sexual behavior by this producer towards associated women. This sparked viral exposure to the #MeToo movement, with several other women in the film industry beginning to stand up to this injustice. This led the treatment of women in not only the film industry, but all industries, to be a greater concern in the public eye. Initiatives were quickly undertaken to boost representation of women in these industries to eliminate this behavior.

Our research question is:

Did the increased exposure to the public of women in the film industry lead to an increase in the ratings of films with female leads?

Data Processing

The data we utilized for this project is acquisitioned from the IMDb (Internet Movie Database) Non-Commercial dataset. This is a comprehensive database consisting of seven individual TSV files that we processed in Python. Although being founded in 1990, IMDb traces back films until around 1898. In total, we have access to just over 11 million observations.

Each of these seven files contained a common label as "*tconst*" or "*title constant*" by which we could merge them. After merging these files, we then began the creation of dummy variables for analysis. A key variable of note is our female dummy "*is_actress*" (created based off lead profession listing "actress"). The explicit gender of leads was not listed, so we used this to infer. With the MeToo movement beginning in 2017, we also created a dummy to represent films that

were released during or after 2017. Dummy variables for adult-rated films and genres were also created. The dataset was then filtered to only include "*movies*" and "*tvmovies*" (for example, shorts and series were removed). We performed a two different analyzes by using the full set, and a subset of films released after the year 2015.

Data Dictionary

- tconst
  - matching process variable
- is_actress
  - dummy variable for female lead
- is_after_2017
  - dummy variable for films after 2017
- is_actress:is_after_2017
  - interaction term, causal impact
- numVotes
  - number of votes for film
- runtimeMinutes
  - length of film in mimutes
- isAdult
  - dummy for adult rated films
- is_action
  - dummy for action genre
- is_drama
  - dummy for drama genre
- is_romance
  - dummy for romance genre
- is_scifi
  - dummy for scifi genre
- is_comedy
  - dummy for comedy genre
- is_adventure
  - dummy for adventure genere
- is_documentary
  - dummy for documentary genre
- is_fantasy
  - dummy for fantasy genre
- is_history
  - dummy for history genre
- is_horror
  - dummy for horror genre
- is_biography
  - dummy for biography genre

Methodology

We employed the Difference-in-Differences (DiD) strategy to analyze our data.

Notated as:

$$averageRating = \alpha + \beta D_{isactress} + \gamma X_{after2017} + \delta_{DD}(D_{isactress} * X_{after2017}) + \theta controls + \varepsilon$$

Below is a table outlining each variable's coefficient and meaning.

	Pre	**Post**	**Diff (Columns)**
Not Treated	α	$\alpha + \gamma$	γ
Treated	$\alpha + \beta$	$\alpha + \beta + \gamma + \delta_{DD}$	$\gamma + \delta_{DD}$
Diff (Rows)	β	$\beta + \delta_{DD}$	δ_{DD}

Difference-in-Differences is an applicable approach to our question because the first difference removes any time-invariant differences (dependent on individuals), and the second difference removes time-dependent differences (relevant shocks in period). After taking both differences we are left with the coefficient on the interaction term (δ_{DD}). This is our variable of interest and shows us our true causal impact.

The key assumption behind the DiD specification is the concept of parallel trends. Parallel trends implies that the two groups (for at least one to two periods before treatment) behaved the same or similarly. This implies that without treatment, both groups would have had the same or similar outcomes. This way we can be certain that any difference between the two groups is based on treatment, and not other factors. Parallel trends can be analyzed visually via a graph of the two

groups, or by testing the interaction term variable on the on the periods before treatment. If this result is insignificant, then the parallel trends empirically hold true.

In practice, we used movies without female leads (*is_actress = 0*) as our control, and movies with female leads (*is_actress = 1)* as our treated group. Time is our running variable, with analysis centered around 2017 (popularization of the MeToo movement).

Results

Test of Parallel Trends

For the test of parallel trends, we shifted our dataset to only focus on years 2015 and 2016. Movies made in 2016 were given a dummy variable of "*after"*. Below is the empirical test of these trends.

OLS Regression Results

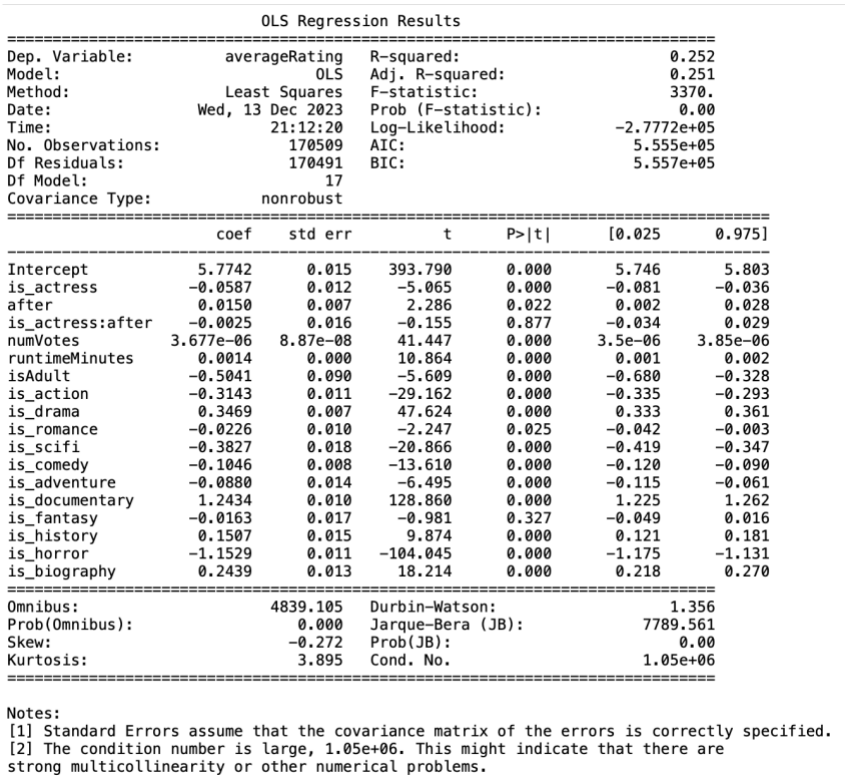
Dep. Variable:	averageRating	R-squared:	0.006
Model:	OLS	Adj. R-squared:	0.006
Method:	Least Squares	F-statistic:	362.0
Date:	Wed, 13 Dec 2023	Prob (F-statistic):	1.97e-234
Time:	21:12:46	Log-Likelihood:	-3.3178e+05
No. Observations:	186075	AIC:	6.636e+05
Df Residuals:	186071	BIC:	6.636e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.2687	0.005	1201.657	0.000	6.258	6.279
is_actress	-0.3003	0.013	-23.461	0.000	-0.325	-0.275
after	0.0068	0.007	0.932	0.351	-0.008	0.021
is_actress:after	0.0107	0.018	0.600	0.549	-0.024	0.046

Omnibus:	5754.022	Durbin-Watson:	1.316
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6307.020
Skew:	-0.441	Prob(JB):	0.00
Kurtosis:	3.186	Cond. No.	7.32

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

As can be seen, our interaction variable is insignificant. Worth noting is the coefficient on *"is_actress",* and that it is negative. This implies that films with female leads receive lower ratings. Furthermore, we tested our interaction term with controls to see if this impacted the insignificance.



OLS Regression Results

Dep. Variable:	averageRating	R-squared:	0.252
Model:	OLS	Adj. R-squared:	0.251
Method:	Least Squares	F-statistic:	3370.
Date:	Wed, 13 Dec 2023	Prob (F-statistic):	0.00
Time:	21:12:20	Log-Likelihood:	-2.7772e+05
No. Observations:	170509	AIC:	5.555e+05
Df Residuals:	170491	BIC:	5.557e+05
Df Model:	17		
Covariance Type:	nonrobust		

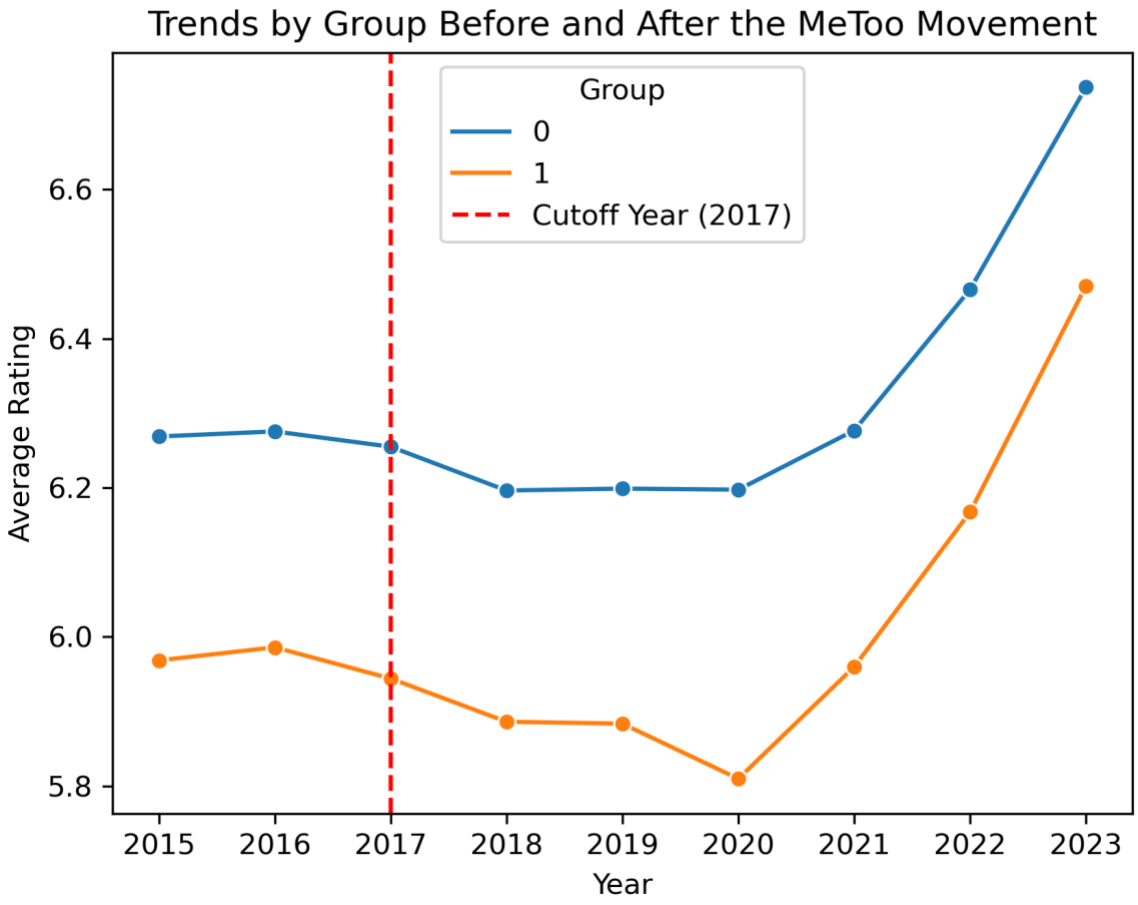
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.7742	0.015	393.790	0.000	5.746	5.803
is_actress	-0.0587	0.012	-5.065	0.000	-0.081	-0.036
after	0.0150	0.007	2.286	0.022	0.002	0.028
is_actress:after	-0.0025	0.016	-0.155	0.877	-0.034	0.029
numVotes	3.677e-06	8.87e-08	41.447	0.000	3.5e-06	3.85e-06
runtimeMinutes	0.0014	0.000	10.864	0.000	0.001	0.002
isAdult	-0.5041	0.090	-5.609	0.000	-0.680	-0.328
is_action	-0.3143	0.011	-29.162	0.000	-0.335	-0.293
is_drama	0.3469	0.007	47.624	0.000	0.333	0.361
is_romance	-0.0226	0.010	-2.247	0.025	-0.042	-0.003
is_scifi	-0.3827	0.018	-20.866	0.000	-0.419	-0.347
is_comedy	-0.1046	0.008	-13.610	0.000	-0.120	-0.090
is_adventure	-0.0880	0.014	-6.495	0.000	-0.115	-0.061
is_documentary	1.2434	0.010	128.860	0.000	1.225	1.262
is_fantasy	-0.0163	0.017	-0.981	0.327	-0.049	0.016
is_history	0.1507	0.015	9.874	0.000	0.121	0.181
is_horror	-1.1529	0.011	-104.045	0.000	-1.175	-1.131
is_biography	0.2439	0.013	18.214	0.000	0.218	0.270

Omnibus:	4839.105	Durbin-Watson:	1.356
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7789.561
Skew:	-0.272	Prob(JB):	0.00
Kurtosis:	3.895	Cond. No.	1.05e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.05e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Still, "*is_actress:after"* is insignificant and maintains a negative coefficient.

Testing the Full Sample



The figure above graphically illustrates ours parallel trends from the pre-treatment period, as well as showing our treatment point (2017) and the trends following. The blue line represents films without female leads (*is_actress* = 0), and the orange represents films with female leads (*is_actress* = 1). Notably, the parallel trends continue until 2020. It is at this time where films with female leads saw a sharp decrease in ratings, before recovering to levels equal to 2015.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          averageRating   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     3785.
Date:                Wed, 13 Dec 2023   Prob (F-statistic):               0.00
Time:                        16:05:23   Log-Likelihood:            -5.3463e+06
No. Observations:             3115684   AIC:                         1.069e+07
Df Residuals:                 3115680   BIC:                         1.069e+07
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       6.2017      0.001   6760.454      0.000       6.200       6.203
is_actress                     -0.1613      0.002    -73.143      0.000      -0.166      -0.157
is_after_2017                   0.1189      0.002     52.455      0.000       0.114       0.123
is_actress:is_after_2017       -0.1506      0.005    -27.969      0.000      -0.161      -0.140
==============================================================================
Omnibus:                    76437.475   Durbin-Watson:                   1.204
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            83965.865
Skew:                          -0.371   Prob(JB):                         0.00
Kurtosis:                       3.311   Cond. No.                         7.55
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The full sample size contained over 3 million observations. The table above shows our treatment variable, *is_actress*, our time variable *is_after_2017,* and our interaction *is_actress:is_after_2017*. All variables were statistically significant, still showing negative coefficients for films with female leads, and those after 2017.

OLS Regression Results

Dep. Variable:	averageRating	R-squared:	0.184
Model:	OLS	Adj. R-squared:	0.184
Method:	Least Squares	F-statistic:	3.723e+04
Date:	Wed, 13 Dec 2023	Prob (F-statistic):	0.00
Time:	16:26:23	Log-Likelihood:	-4.4897e+06
No. Observations:	2812462	AIC:	8.979e+06
Df Residuals:	2812444	BIC:	8.980e+06
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.9819	0.002	3121.803	0.000	5.978	5.986
is_actress	-0.0470	0.002	-22.488	0.000	-0.051	-0.043
is_after_2017	0.0063	0.002	2.961	0.003	0.002	0.010
is_actress:is_after_2017	-0.0512	0.005	-10.117	0.000	-0.061	-0.041
numVotes	3.313e-06	1.96e-08	169.086	0.000	3.27e-06	3.35e-06
runtimeMinutes	0.0003	8.68e-06	40.249	0.000	0.000	0.000
isAdult	-0.4940	0.007	-75.216	0.000	-0.507	-0.481
is_action	-0.3859	0.003	-153.422	0.000	-0.391	-0.381
is_drama	0.2999	0.002	175.416	0.000	0.297	0.303
is_romance	-0.0374	0.002	-16.448	0.000	-0.042	-0.033
is_scifi	-0.4972	0.005	-106.487	0.000	-0.506	-0.488
is_comedy	-0.0857	0.002	-47.225	0.000	-0.089	-0.082
is_adventure	-0.1392	0.003	-45.489	0.000	-0.145	-0.133
is_documentary	1.1366	0.003	451.432	0.000	1.132	1.142
is_fantasy	0.0315	0.004	7.689	0.000	0.023	0.039
is_history	0.2154	0.004	50.640	0.000	0.207	0.224
is_horror	-1.0248	0.003	-342.595	0.000	-1.031	-1.019
is_biography	0.2004	0.004	49.735	0.000	0.193	0.208

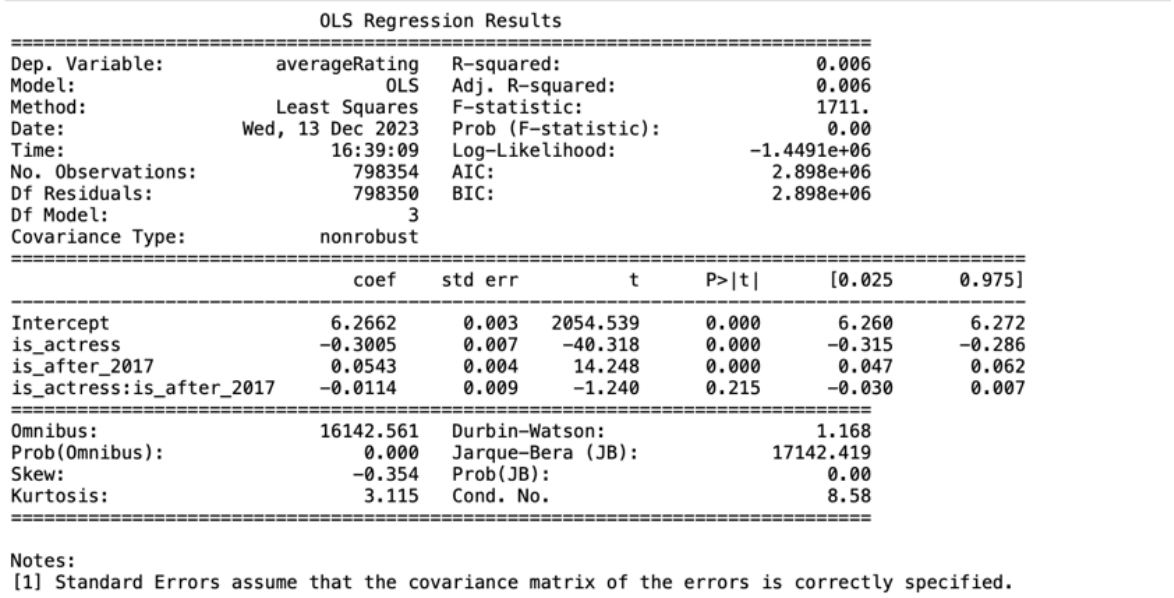
Omnibus:	84946.867	Durbin-Watson:	1.288
Prob(Omnibus):	0.000	Jarque-Bera (JB):	135868.315
Skew:	-0.288	Prob(JB):	0.00
Kurtosis:	3.909	Cond. No.	3.42e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.42e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Upon adding controls, our sample size decreased to 2.8 million observations. All variables remained statistically significant, including our variable of interest *is_actress:is_after_2017*.

Subsetting Sample: 2015-2023

To condense our analysis and make it clearer, we created a separate data frame in Python to include only movies from 2015 to 2023. This brought our observations to just shy of 800,000. Both actress and interaction, without controls, remained negative. However, the interaction term was not statistically significant. This could possibly be due to the decrease in sample size (from 2.8 million to 800,000), which would have skewed the distribution of results.



OLS Regression Results

Dep. Variable:	averageRating	R-squared:	0.006
Model:	OLS	Adj. R-squared:	0.006
Method:	Least Squares	F-statistic:	1711.
Date:	Wed, 13 Dec 2023	Prob (F-statistic):	0.00
Time:	16:39:09	Log-Likelihood:	-1.4491e+06
No. Observations:	798354	AIC:	2.898e+06
Df Residuals:	798350	BIC:	2.898e+06
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.2662	0.003	2054.539	0.000	6.260	6.272
is_actress	-0.3005	0.007	-40.318	0.000	-0.315	-0.286
is_after_2017	0.0543	0.004	14.248	0.000	0.047	0.062
is_actress:is_after_2017	-0.0114	0.009	-1.240	0.215	-0.030	0.007

Omnibus:	16142.561	Durbin-Watson:	1.168
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17142.419
Skew:	-0.354	Prob(JB):	0.00
Kurtosis:	3.115	Cond. No.	8.58

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Upon adding controls below, observations became almost 725,000. Our interaction variable became just slightly insignificant at this point, likely due to the decrease we see in 2020. Worth noting is that only 3 of our explanatory variables were statistically insignificant, with almost every genre showing a slight decrease in ratings. What we did not account for was films that had multiple genres listed.

For instance, some films could have been listed as “documentary, history, action” as a singular genre category. To prevent multicollinearity issues as far as the dummy variables are concerned, we went off the first genre listed. In this example, the film would be listed as “*is_documentary”*.

OLS Regression Results

Dep. Variable:	averageRating	R-squared:	0.229
Model:	OLS	Adj. R-squared:	0.229
Method:	Least Squares	F-statistic:	1.268e+04
Date:	Wed, 13 Dec 2023	Prob (F-statistic):	0.00
Time:	16:39:14	Log-Likelihood:	-1.2074e+06
No. Observations:	724165	AIC:	2.415e+06
Df Residuals:	724147	BIC:	2.415e+06
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.9076	0.005	1278.196	0.000	5.899	5.917
is_actress	-0.0692	0.007	-10.218	0.000	-0.083	-0.056
is_after_2017	0.0703	0.003	20.381	0.000	0.064	0.077
is_actress:is_after_2017	-0.0162	0.008	-1.936	0.053	-0.033	0.000
numVotes	4.097e-06	5.41e-08	75.766	0.000	3.99e-06	4.2e-06
runtimeMinutes	0.0001	1.41e-05	8.879	0.000	9.76e-05	0.000
isAdult	0.0277	0.061	0.453	0.650	-0.092	0.148
is_action	-0.3052	0.006	-54.944	0.000	-0.316	-0.294
is_drama	0.3581	0.004	97.029	0.000	0.351	0.365
is_romance	-0.0118	0.005	-2.267	0.023	-0.022	-0.002
is_scifi	-0.5366	0.009	-57.643	0.000	-0.555	-0.518
is_comedy	-0.1140	0.004	-28.953	0.000	-0.122	-0.106
is_adventure	-0.0561	0.007	-7.811	0.000	-0.070	-0.042
is_documentary	1.2084	0.005	251.139	0.000	1.199	1.218
is_fantasy	-0.0165	0.009	-1.907	0.057	-0.033	0.000
is_history	0.1265	0.008	14.944	0.000	0.110	0.143
is_horror	-1.1624	0.006	-211.161	0.000	-1.173	-1.152
is_biography	0.1708	0.007	23.065	0.000	0.156	0.185

Omnibus:	10796.865	Durbin-Watson:	1.243
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18054.257
Skew:	-0.127	Prob(JB):	0.00
Kurtosis:	3.731	Cond. No.	1.16e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.16e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Conclusion

We conclude that films with a female leads throughout the full sample receive about 0.16 less stars on average without controls, and 0.04 less with controls. When subset to only years 2015-2023, films with female leads saw a decrease in ratings of 0.3 stars without controls, and 0.07 with controls.

Our variable of causal impact (apart from a 0.053 p value for 2015-2023) remained statistically significant throughout the entirety of our analysis, showing that female films on average after 2017 saw a decrease in their ratings. Our original viewpoint was that these films would see an increase in ratings, and, while marginal, not see a decrease.

The dip in 2020 deems worthy of future analysis to divulge why women-led films saw a sharp decrease. This could be due to an impact of Covid-19 on films with female leads and would require further research.

Works Cited

Brittain, A. (2023, November 17). *Me Too Movement*. Retrieved December 2023, from Britannica website: https://www.britannica.com/topic/Me-Too-movement

Burke, T. (n.d.). *History & Inception*. Retrieved December 2023, from MeToo Movement website: https://metoomvmt.org

IMDb. (n.d.). *IMDb Non-Commercial Datasets*. Retrieved December 2023, from IMDb Developer website: https://developer.imdb.com/non-commercial-datasets/