

# Predicting Horror IMDB Scores of Female Audiences

Jillian Etheredge



# 680 Horror Movies

That's a lot of movies

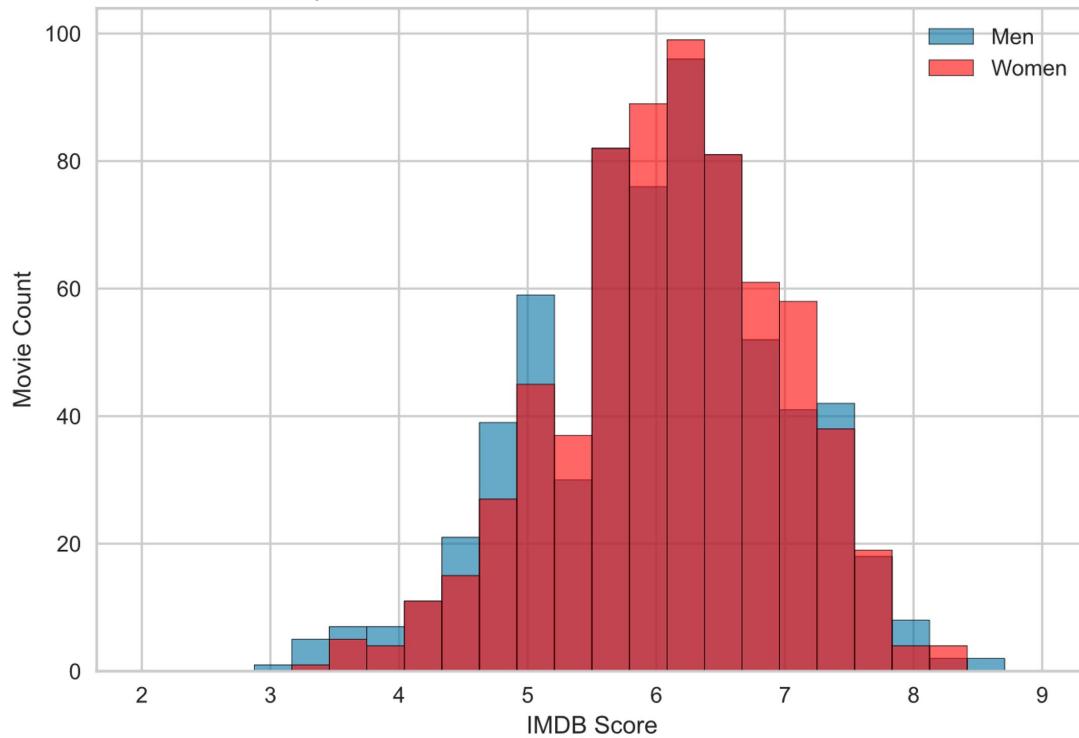
# 43,612,742 Votes

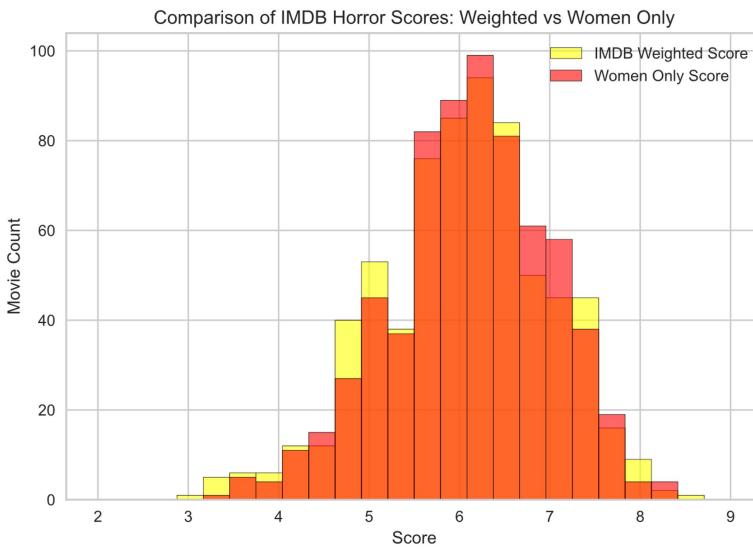
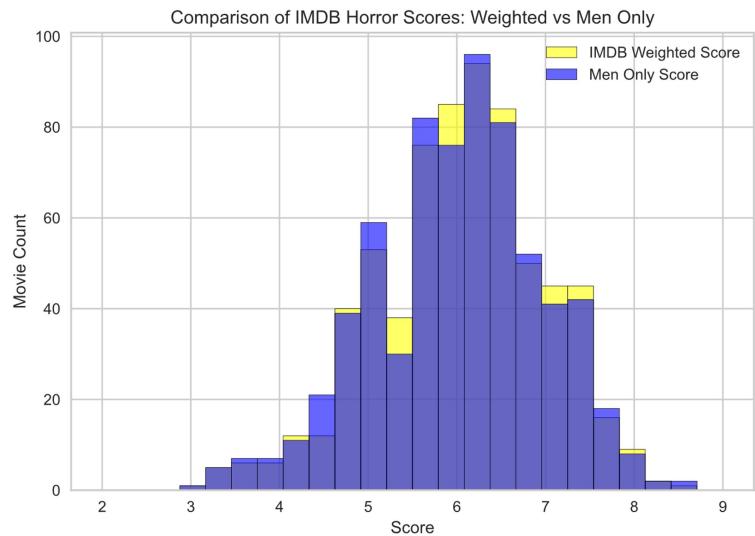
And a lot of votes

# Only 17.5% Women's Votes

The true horror is the lack of women

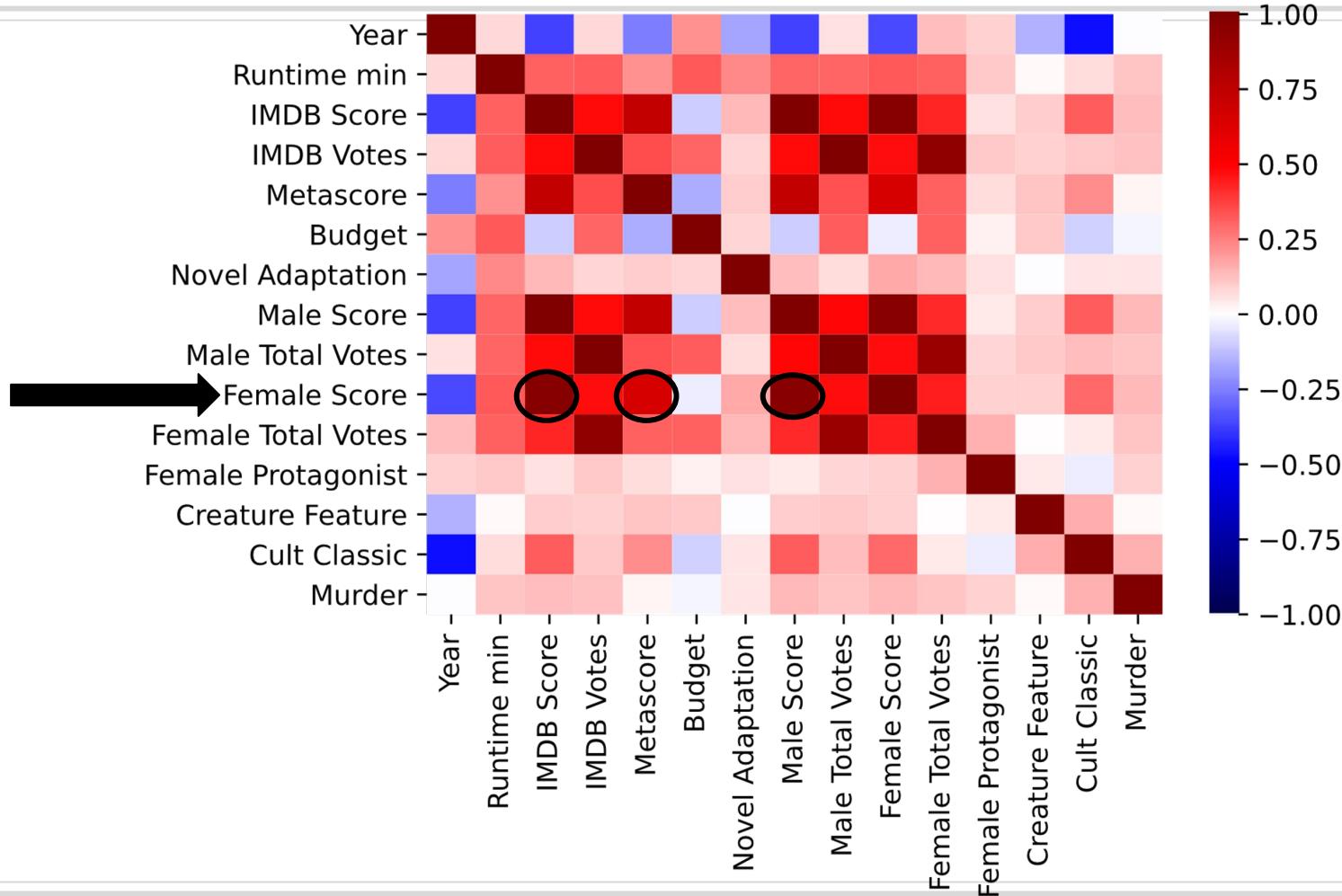
Comparison of IMDB Horror Scores for Men and Women





# Methodology

- ❑ Webscraping IMDB with BeautifulSoup
- ❑ Find correlations with target
- ❑ Feature selection based on potential use cases:
  - Previously released movies
  - Movies pending release
  - Pitching a new movie
- ❑ Linear Regression Modeling



# *Features*

## *Numeric*

- Year
- Runtime (minutes)
- Male Score
- Metascore
- Budget

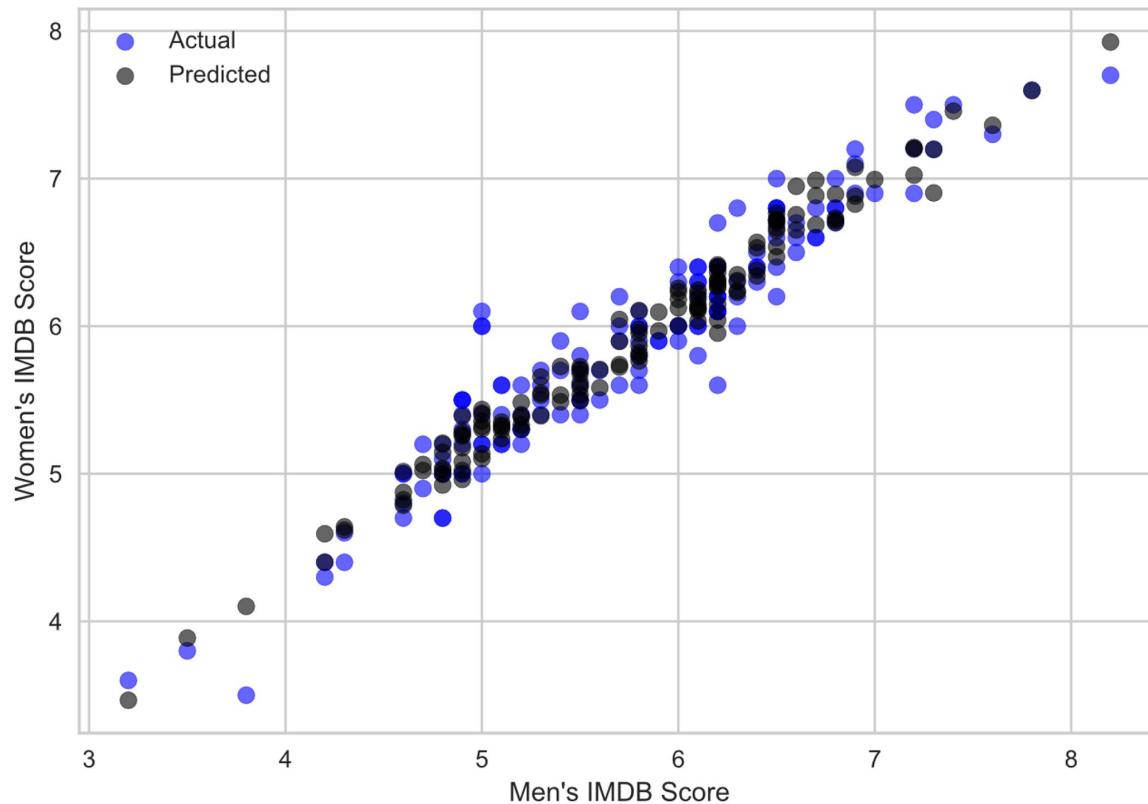
## *Categorical*

- Director
- Genre
- Novel Adaptation
- Plot tags:
  - Monster
  - Cult Classic
  - Female Protagonist
  - Murder

# *Predicting Score for Previously Released Movies*

- Most predictive model
- Includes Male Score and Metascore
- Using Ridge Regression:
  - $R^2$ : 0.926
- Highest Lasso Regression Coef:
  - Musical: 0.541
  - Male Score: 1.128

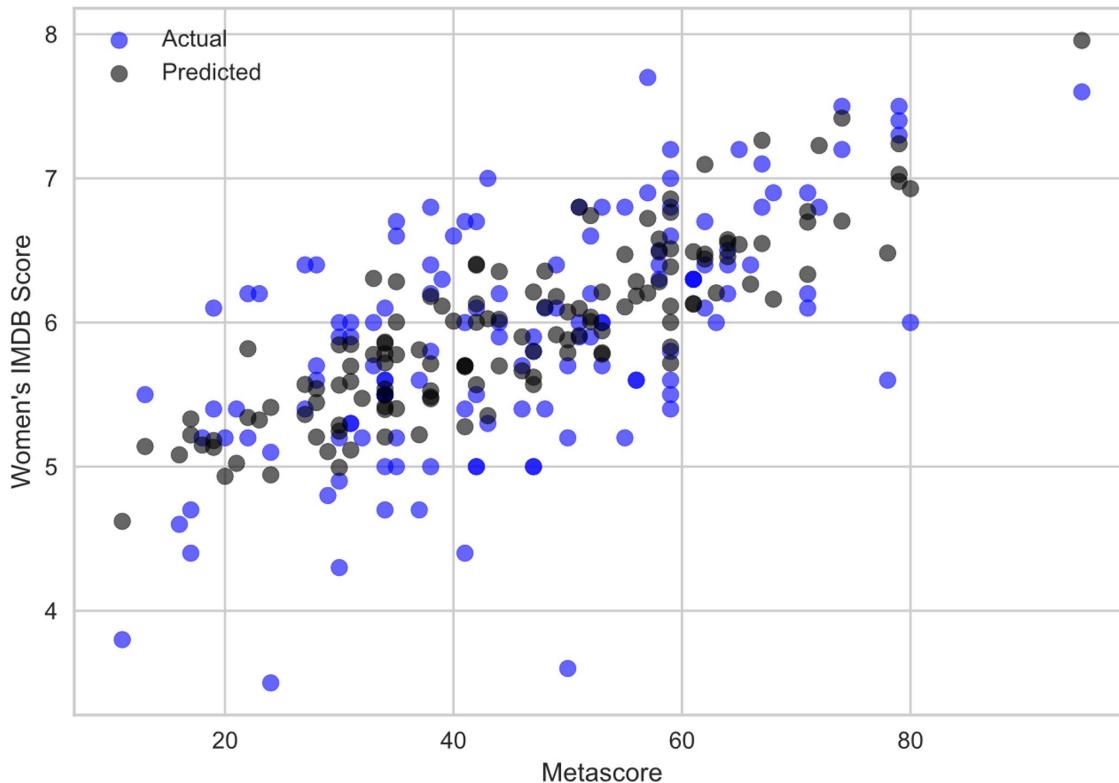
Actual vs Predicted Test Values



# *Predicting Score for Movies Pending Release*

- Second most predictive model
- Includes Metascore
- Using Ridge Regression:
  - $R^2$ : 0.516
- Highest Lasso Regression Coef:
  - Director James Wan: 0.768
  - Metascore: 0.734

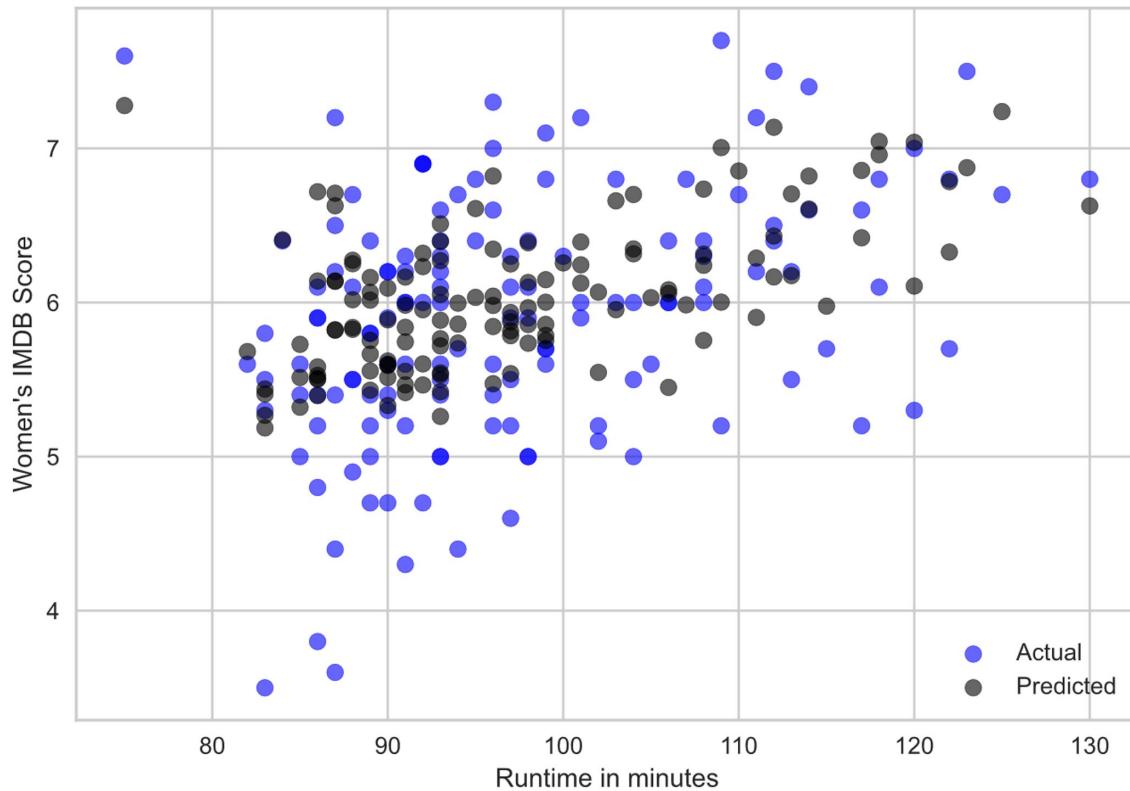
Actual vs Predicted Test Values



# *Predicting Score for Movie Pitch*

- Least predictive model
- Does not include audience or critic scores
- Using Ridge Regression:
  - $R^2$ : 0.330
- Highest Lasso Regression Coef:
  - Director James Wan: 0.742
  - Runtime: 0.322

Actual vs Predicted Test Values



# *Conclusions*

- The IMDB data I scraped was insufficient to predict the scores of a female audience without audience or critic features
- Movies that do well with a male audience and critics typically do well with female audiences
- Directors often had a higher correlation with scoring success than plot tags

# *Future Work*

- Collect more data
- Add new features
- Narrow critic review scoring to female critics
- Test with different model types

*thanks!*

Any questions?

*Here's the appendix*

## Breakdown on Male and Female Statistics

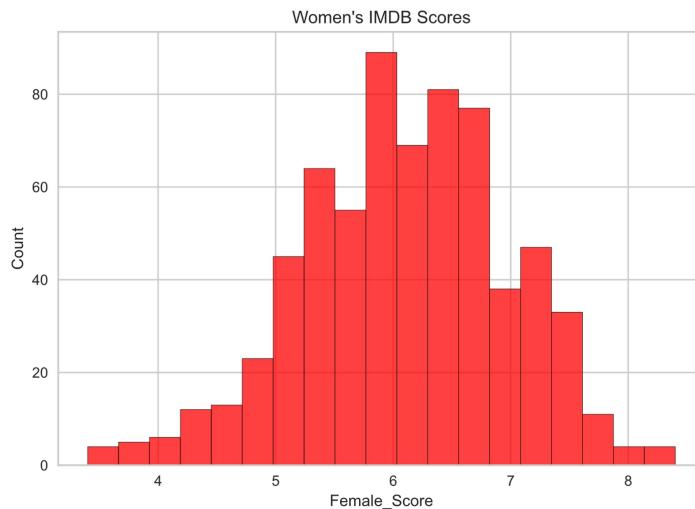
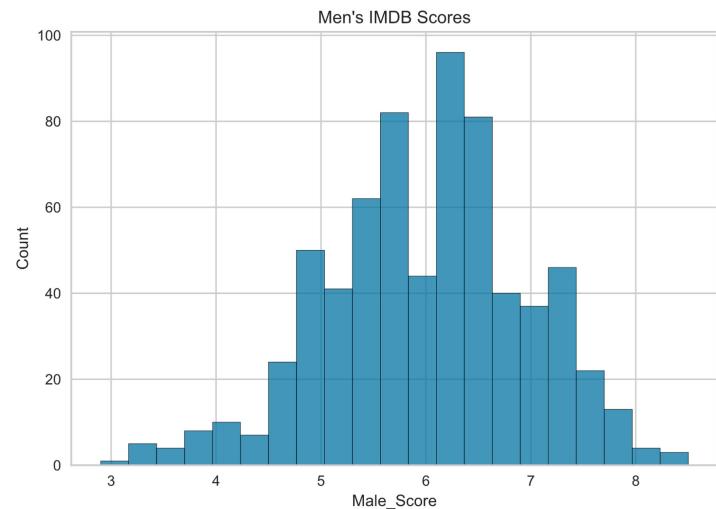
Looking at 680 horror films  
43,612,742 Votes

### *Men*

- 82.5% of total votes
- Least votes: 3,105
- Most votes: 530,475

### *Women*

- 17.5% of total votes
- Least votes: 638
- Most votes: 124,287

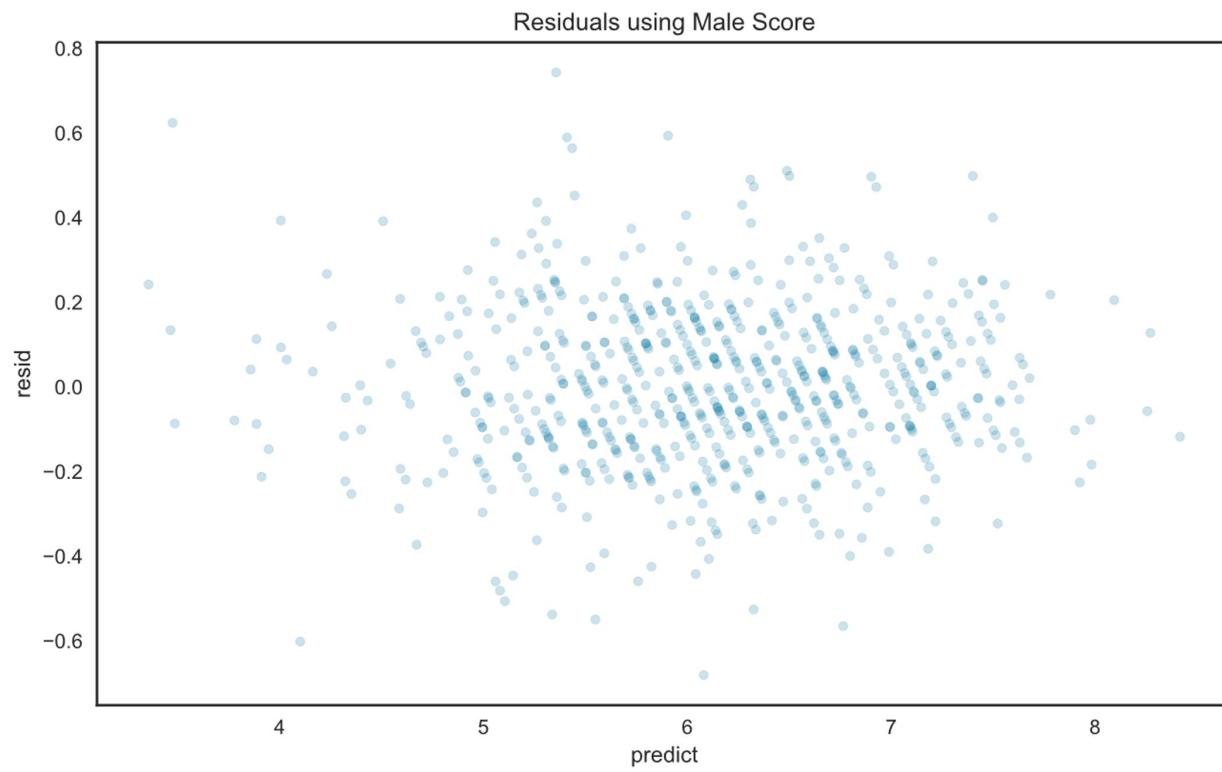


# All Lasso Coefficients for Male Score

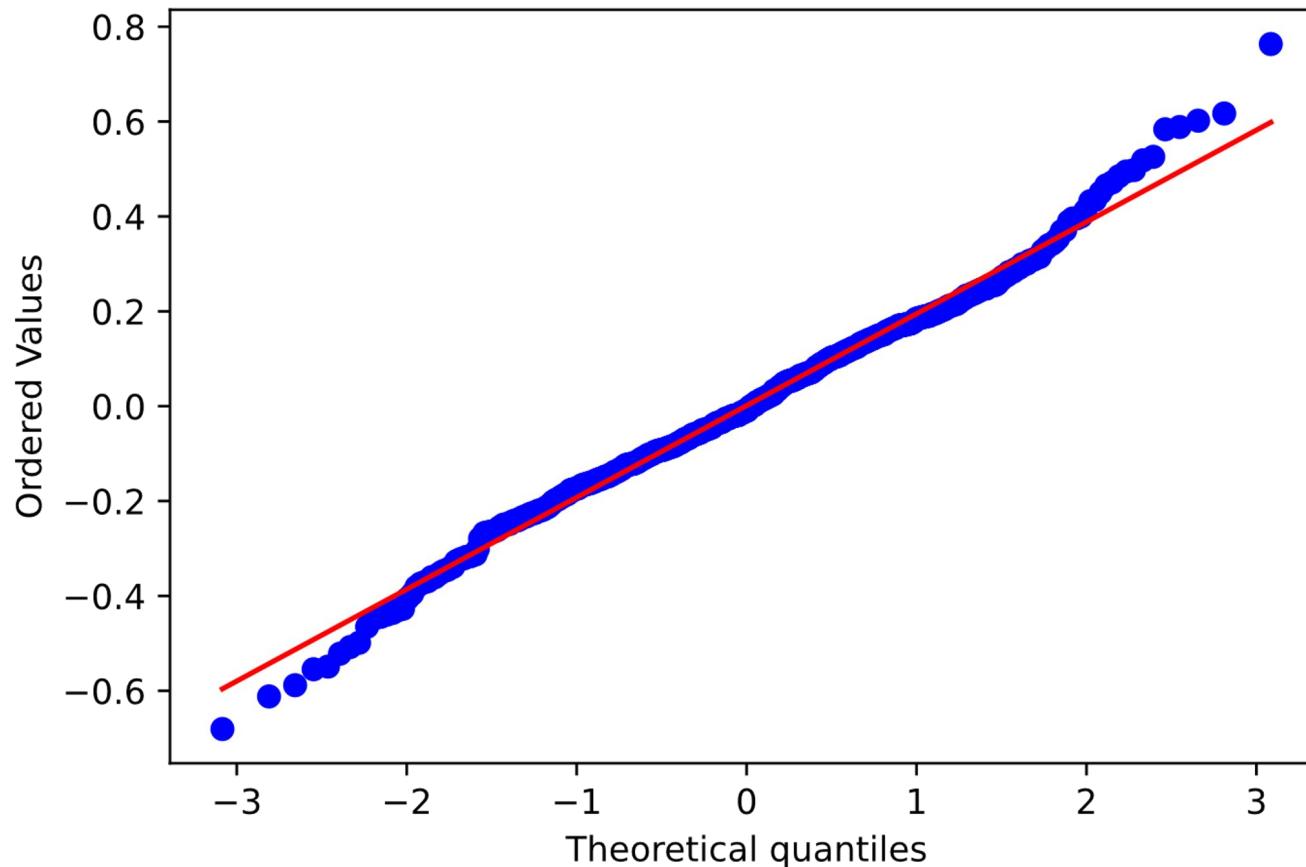
'Comedy', 0.00374410565170195),  
('Thriller', 0.004771514908417107),  
('Murder', 0.008526083346423472),  
('Director\_Other', 0.013393612405557413),  
('Mystery', -0.01481949775418862),  
(Action', 0.015546079956282256),  
(Animation', 0.01791539548205807),  
(Sci-Fi', -0.018447154828122606),  
(Runtime\_min', 0.019136440852901348),  
(Creature\_Feature', -0.025990712794288103),  
(Director\_Christopher Landon', 0.029258912742368498),  
(Year', -0.033901339203013435),  
(Cult\_Classic', -0.03515161445435634),  
(Crime', 0.038849364450716704),  
(Director\_James Wan', 0.053639977181636926),  
(Drama', 0.05870140987944806),  
(Novel\_Adaptation', 0.06485839612701477),  
(Female\_Protagonist', 0.06508444462579362),  
(Female\_Protagonist', 0.06508444462579362),  
(Director\_John Carpenter', -0.06519478594882523),  
(Director\_Wes Craven', 0.06909392025773875),  
(Director\_Guillermo del Toro', 0.06940661932824925),  
(Director\_Darren Lynn Bousman', 0.07154678837645537),  
(Fantasy', 0.0804360048057009),  
(Director\_William Brent Bell', 0.08291856812025027),  
(Romance', 0.09864763918685579),  
(Director\_George A. Romero', -0.10599647025507941),  
(Director\_Sam Raimi', -0.10654228046664076),  
(Biography', -0.11238492169999238),  
(Metascore', -0.13476013151976446),  
(Director\_Tobe Hooper', -0.20775138076098412),  
(Family', 0.20941716655494622),  
(Director\_Alexandre Aja', -0.2727310440675254),  
(Director\_Paul W.S. Anderson', 0.29625441073721337),  
(Musical', 0.5413263228446936),  
(Male\_Score', 1.1280723854762331)

# *Predicting Score for Previously Released Movies*

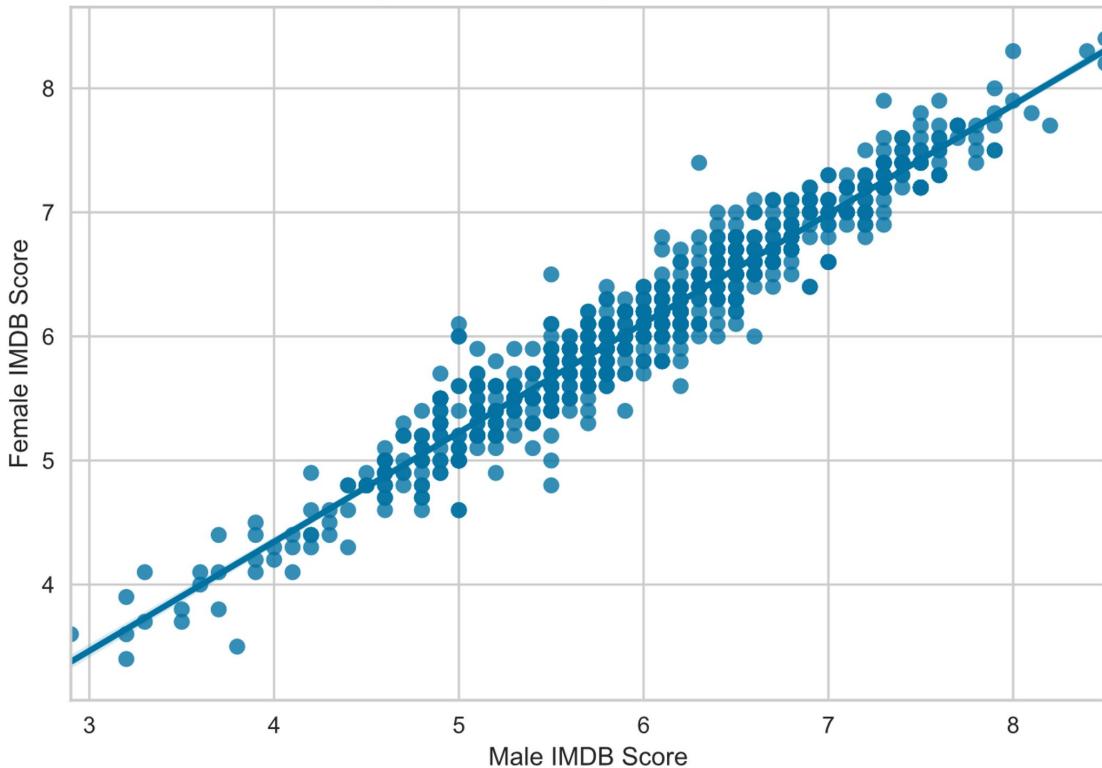
- Most predictive model
- Includes Male Score and Metascore
- Using Ridge Regression:
  - $R^2$ : 0.926
  - Mean Absolute Error: 0.169
  - Mean Squared Error: 0.048
  - Root Mean Squared Error: 0.218
- Highest Lasso Regression Coef:
  - Musical: 0.541
  - Male Score: 1.128



Normal Q-Q plot including Male Score



Regression Plot using Male IMDB Score



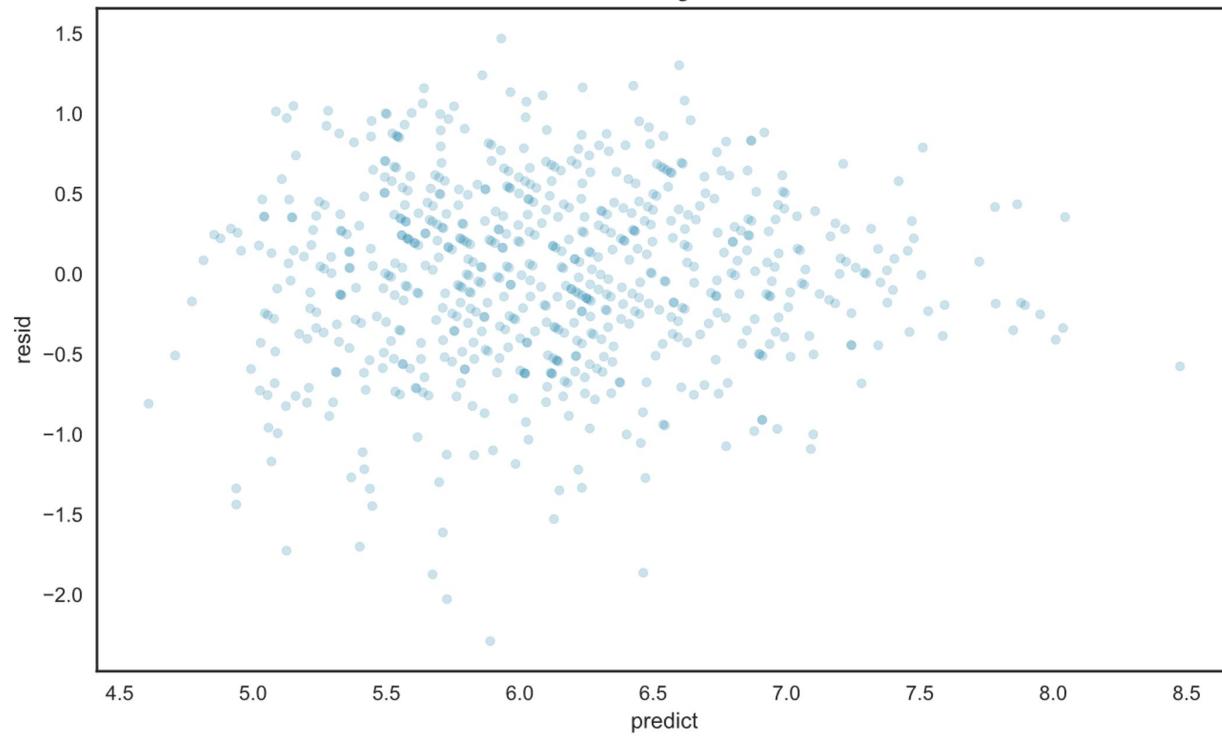
## All Lasso Coefficients for Metascore

('Director\_John Carpenter',  
0.0017838609408893475),  
('Cult\_Classic', 0.018991606439711557),  
('Director\_Other', -0.024057576431460753),  
('Budget', 0.024587684692083896),  
('Drama', 0.02741904293375757),  
('Sci-Fi', -0.045962849242831616),  
('Creature\_Feature', -0.06154201490506926),  
('Thriller', -0.06299565039163672),  
('Fantasy', 0.07559294576516423),  
('Female\_Protagonist', 0.07742710862982138),  
('Mystery', -0.0839206238034574),  
('Musical', 0.1011110759353892),  
('Action', 0.10729240966099617),  
('Adventure', -0.1078219179090507),  
('Runtime\_min', 0.17910877243625914),  
('Crime', 0.18403131024712532),  
('Murder', 0.18997513660319199),  
('Director\_George A. Romero',  
-0.1953199721637179),  
('Director\_Darren Lynn Bousman',  
0.27277029244135514),  
('Year', -0.30896439914855284),  
('Director\_Wes Craven', -0.32992911006449627),  
('Metascore', 0.7338560631923502),  
('Director\_James Wan', 0.7680673769818154)

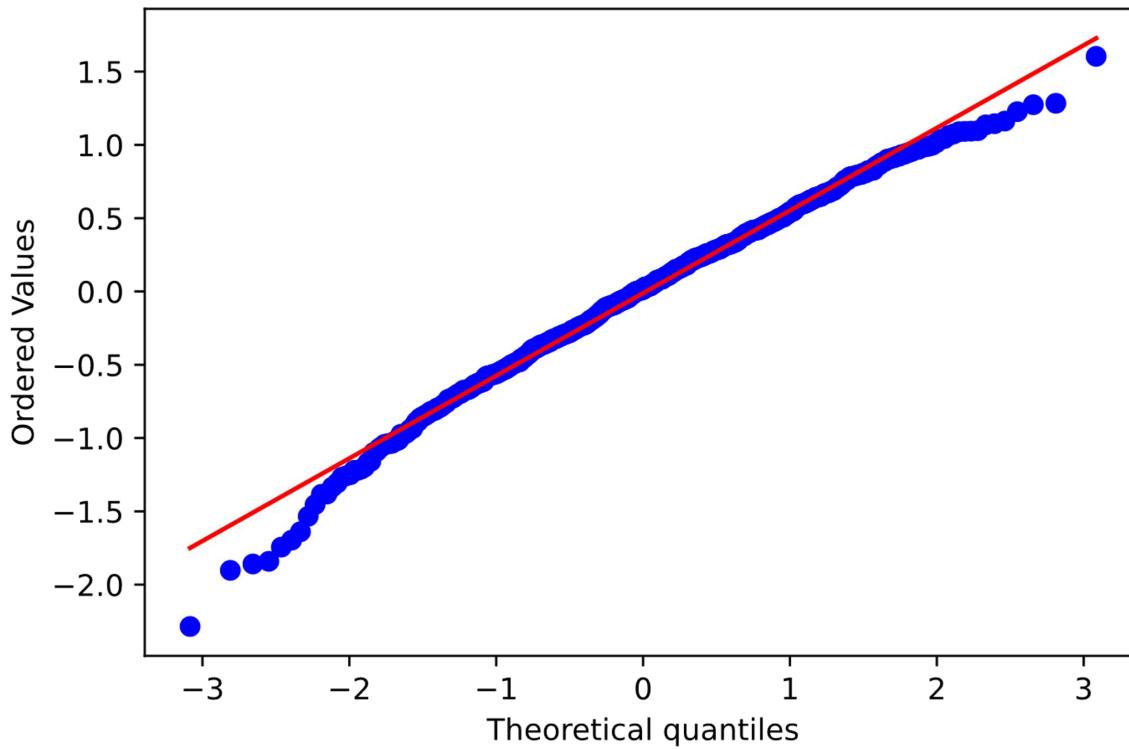
# *Predicting Score for Movies Pending Release*

- Second most predictive model
- Includes Metascore
- Using Ridge Regression:
  - $R^2$ : 0.530
  - Mean Absolute Error: 0.426
  - Mean Squared Error: 0.304
  - Root Mean Squared Error: 0.551
- Highest Lasso Regression Coef:
  - Director James Wan: 0.768
  - Metascore: 0.734

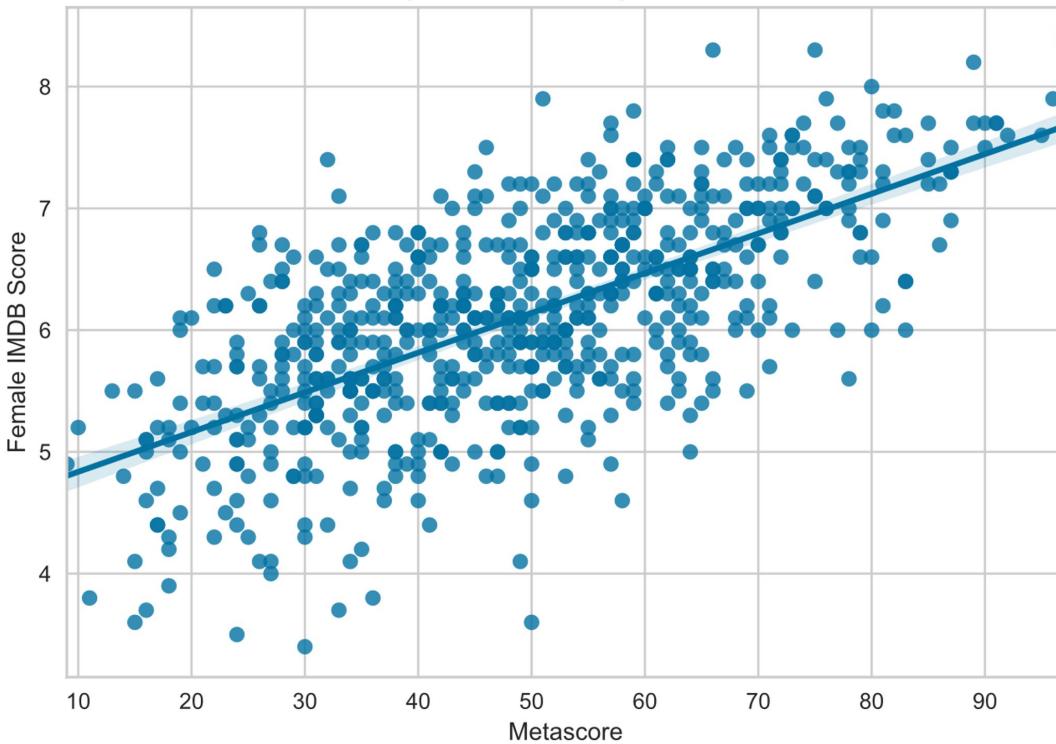
Residuals using Metascore



Normal Q-Q plot including Metascore



Regression Plot using Metascore



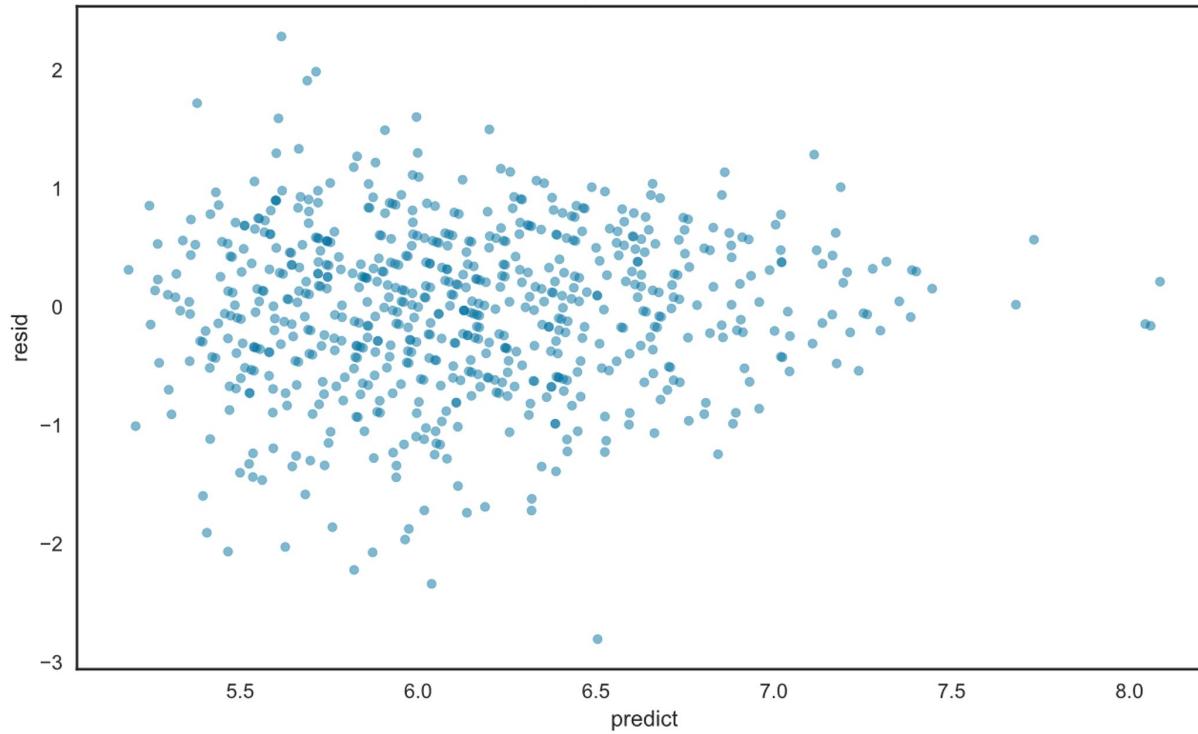
## All Lasso Coefficients for Female Score Predictors

('Director\_Mike Flanagan', 0.017684388867366313), ('Murder', 0.172297996028574),  
('Director\_Christopher Landon', 0.020140161357188345), ('Director\_Other', -0.17363995852834108),  
('Novel\_Adaptation', 0.022799151304034978), ('Adventure', -0.17606758804044692),  
('Fantasy', 0.033696773997248936), ('Female\_Protagonist', 0.20685392425425908),  
('Thriller', -0.039697023358966965), ('Horror', -0.2211880308445127),  
('Sci-Fi', 0.050388893372888426), ('Director\_Tim Burton', 0.23800825465375502),  
('Action', 0.055930292546503654), ('Drama', 0.24088177411667716),  
('Budget', -0.059730197402427826), ('Director\_Rob Zombie', -0.29323922341620456),  
('Mystery', -0.060381835900018645), ('Runtime\_min', 0.3223350267468774),  
('Director\_George A. Romero', -0.06537526623229101), ('Animation', 0.3529213541385356),  
('Director\_Guillermo del Toro', 0.06805521513631949), ('Musical', 0.4448906856631793),  
('Comedy', 0.08570427168757999), ('Year', -0.45624389415144156),  
('Cult\_Classic', 0.09422032727044077), ('Director\_Wes Craven', -0.4681529931909872),  
('Director\_Sam Raimi', 0.10498769612843115), ('Director\_James Wan', 0.742874789797115)  
('Crime', 0.1180425836861281),

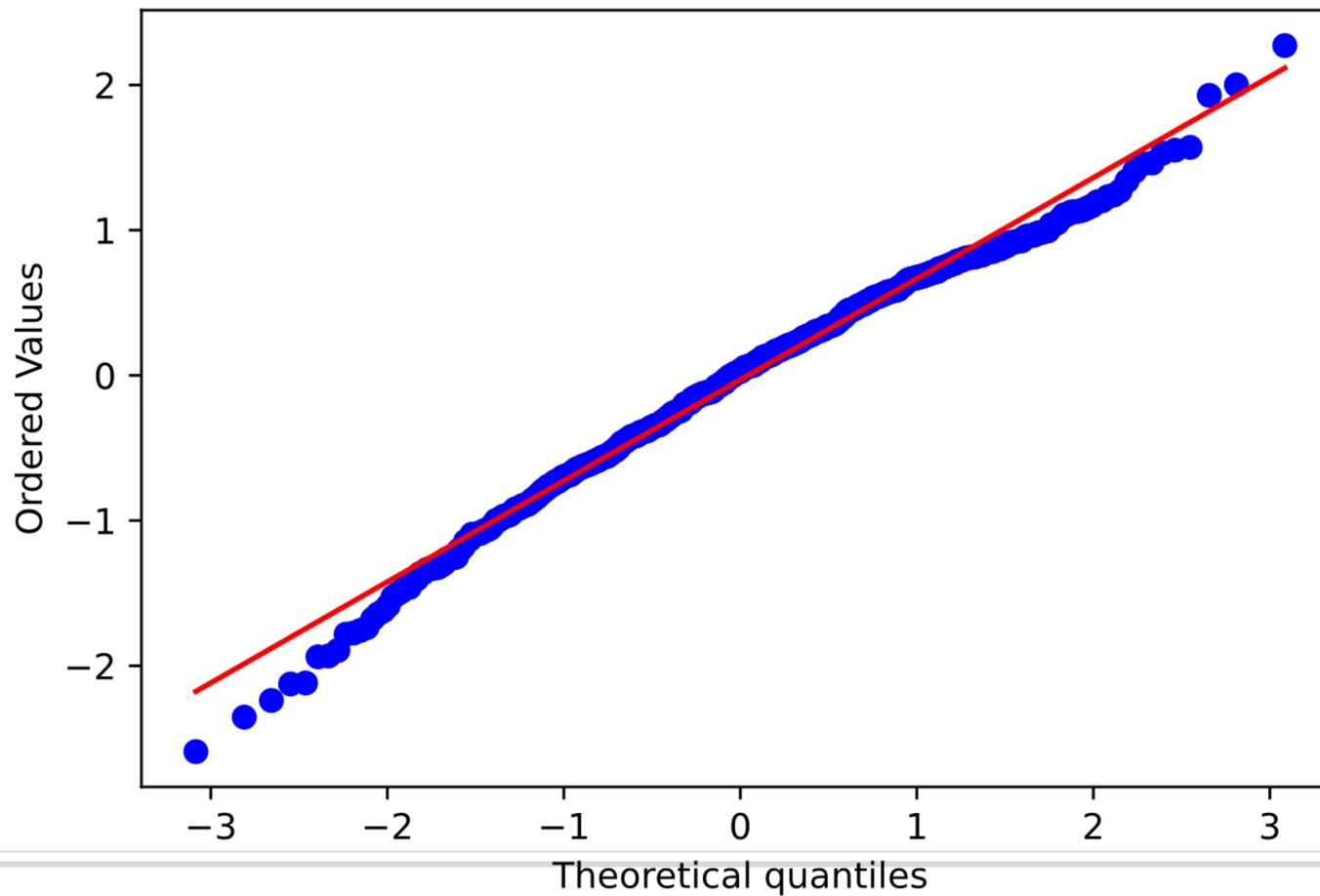
# *Predicting Score for Movie Pitch*

- Least predictive model
- Does not include audience or critic scores
- Using Ridge Regression:
  - $R^2$ : 0.330
  - Mean Absolute Error: 0.510
  - Mean Squared Error: 0.433
  - Root Mean Squared Error: 0.658
- Highest Lasso Regression Coef:
  - Director James Wan: 0.742
  - Runtime: 0.322

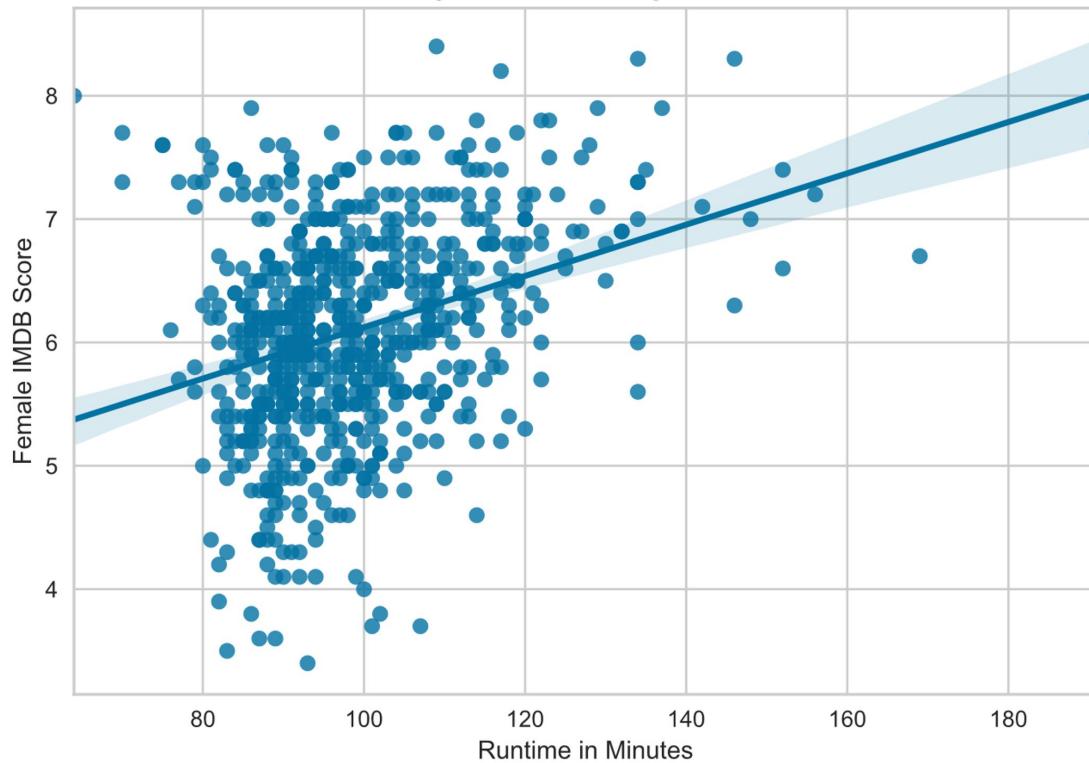
Residuals using Female Score Predictors



### Normal Q-Q plot using Female Score Predictors



Regression Plot using Runtime



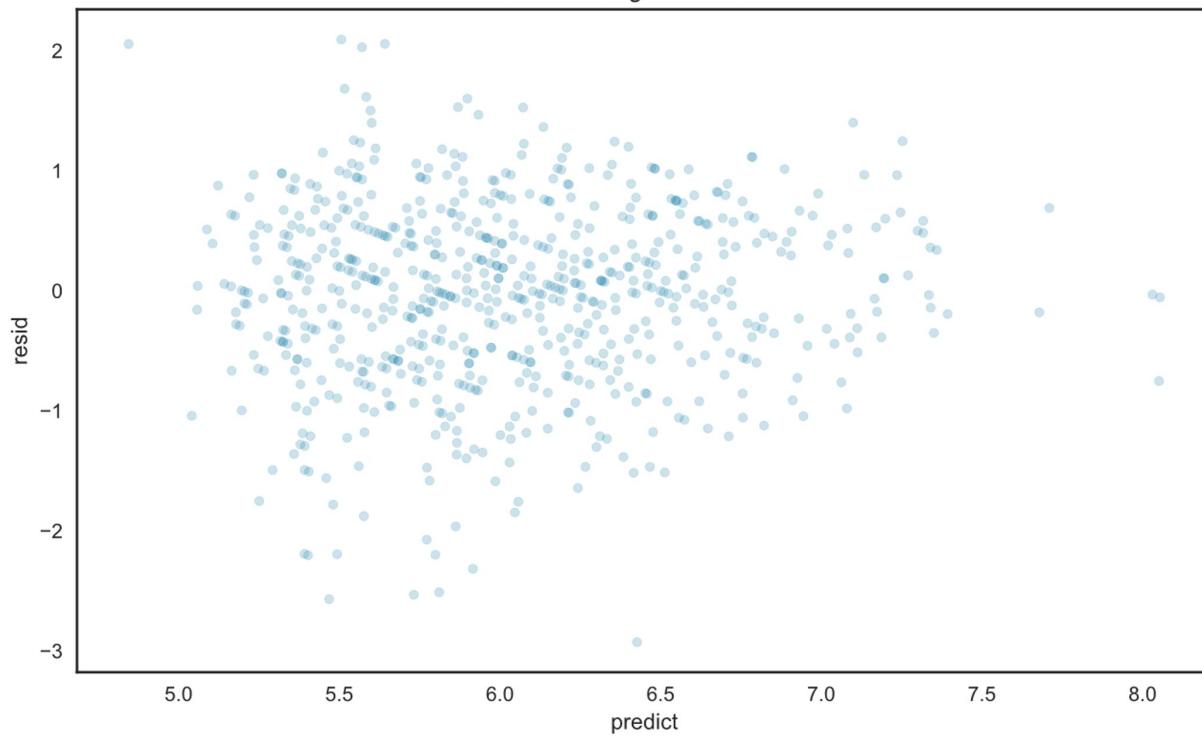
# *Predicting Score for Movie Pitch for male audience*

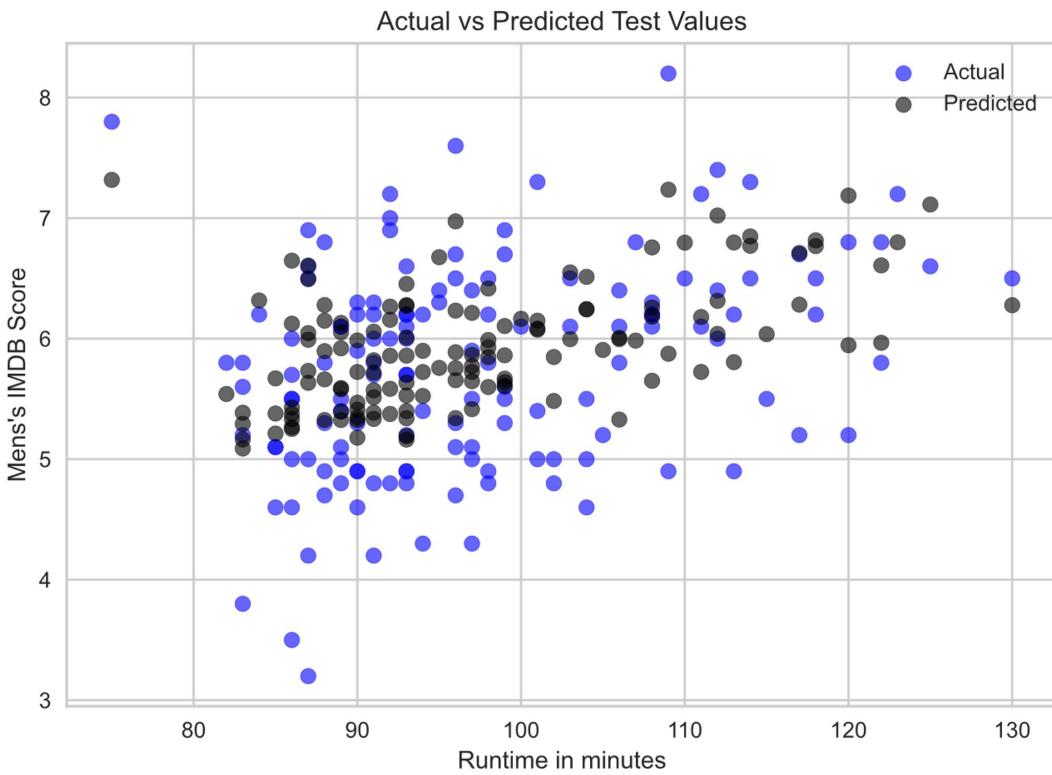
- Least predictive model
- Does not include audience or critic scores
- Using Ridge Regression:
  - $R^2$ : 0.292
  - Mean Absolute Error: 0.588
  - Mean Squared Error: 0.549
  - Root Mean Squared Error: 0.741
- Highest Lasso Regression Coef:
  - Director James Wan: 0.499
  - Runtime: 0.357

## All Lasso Coefficients for Male Score Predictors

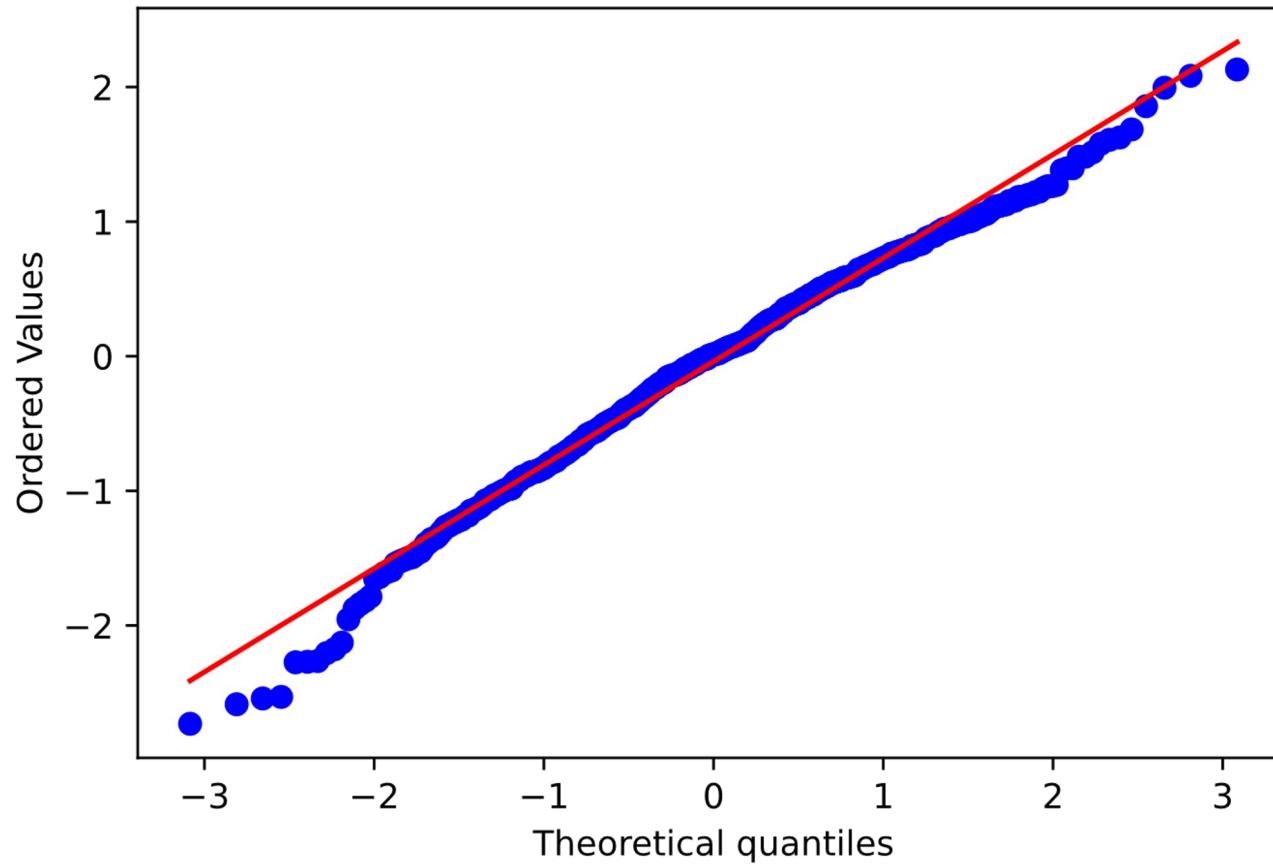
('Director\_David Cronenberg', 0.013163179791430538),  
('Director\_Mike Flanagan', 0.021490328529226863),  
('Mystery', -0.022894196068932098),  
('Thriller', -0.029969965004991425),  
('Creature\_Feature', 0.037678761998297734),  
('Action', 0.05732271898846393),  
('Director\_Guillermo del Toro', 0.06991949625237483),  
('Crime', 0.08216891744084677),  
('Director\_John Carpenter', 0.085841681065091),  
('Comedy', 0.10515232381089208),  
('Budget', -0.11792814002583467),  
('Sci-Fi', 0.12062268185175508),  
('Cult\_Classic', 0.15479692204896695),  
('Female\_Protagonist', 0.1628642691567626),  
('Murder', 0.1631754499044167)  
('Adventure', -0.17662525255856745),  
('Drama', 0.23248704569682294),  
('Director\_Other', -0.23850667330371042),  
('Horror', -0.25304344358777947),  
('Director\_William Brent Bell',  
-0.2878388386961035),  
('Director\_Rob Zombie', -0.3263253033875819),  
('Animation', 0.3371014792062087),  
('Runtime\_min', 0.3588595859325505),  
('Director\_Sam Raimi', 0.36360426868781337),  
('Director\_Tim Burton', 0.4429430064344802),  
('Year', -0.467049560530619),  
('Director\_Wes Craven', -0.6297642065193516),  
('Director\_James Wan', 0.6773651624393858)

Residuals using Male Predictors





### Normal Q-Q plot using Male Score Predictors



Regression Plot using Runtime

