Stat 210 Project

Sofia Silvosa

Introduction

Background and Significance:

There has been recent discussion on adequate female representation in film. Many in the film industry have recently praised the increased representation of women in film. For a while Hollywood produced very little movies with female protagonists; let alone movies with appropriate representation of women (Goodman, 2017). But there has been recent progress in the recent years with an impressive increase in female lead movies in the 2010s; not only that but these films have achieved monetary success. In 2016 with, 29 percent of protagonists in the top 100 box-office hits were women (Goodman, 2017)

However, merely having female protagonists does not paint the full picture if whether a movie actually does a good job in female representation. Many have used the so-called Bechdel test to make this judgement. This test was developed by Allison Bechdel in 1985 and has recently become a digital sensation (O'Meara).

The Bechdel test is a simple test which deems that a movie has adequate female character in their film if it passes the following three criteria: (1) it has to have at least two women in it, who (2) who talk to each other, about (3) something besides a man (https://bechdeltest.com/) .

Many have argued that there has been significant increase of female representation throughout the decades thanks to increased diversity initiatives and more women at the helm of the film industry (UCLA-Hollywood-Diversity-Report-2022-Film). If this the case, then we would expect more recent films, specially those of the 2000s and 2010s, to have a greater amount of films that pass the Bechdel test.

Research Question & Hypothesis

We are interested in evaluating what variables are important in predicting whether a movie form this specific data set passes or fails the Bechdel test. We are interested which variables in our model can help us predict whether a model passes or fails the Bechdel test. We are specifically interested in evaluating whether the budget of the film and teh time period when it was released plays a significnat role since these factors have been remarked as highly important for our

In other words, Does the decade a movie was released predict whether it passes the Bechdel test while controlling for other pivotal variables in our model?

As shiwn by recent news and media, more modern movies seem to have more increased rprpensstaion of women in their films. Thus, we predict that the time period in whcih a movie was released will play a significnat role in whether the movie fails or passes the Bechdel test. More specicially, we predict that movies released in the 200's and 2010's have a greater percentage of movies that pass the Bechdel rather than those that came out in the 20th century represented in our data set.

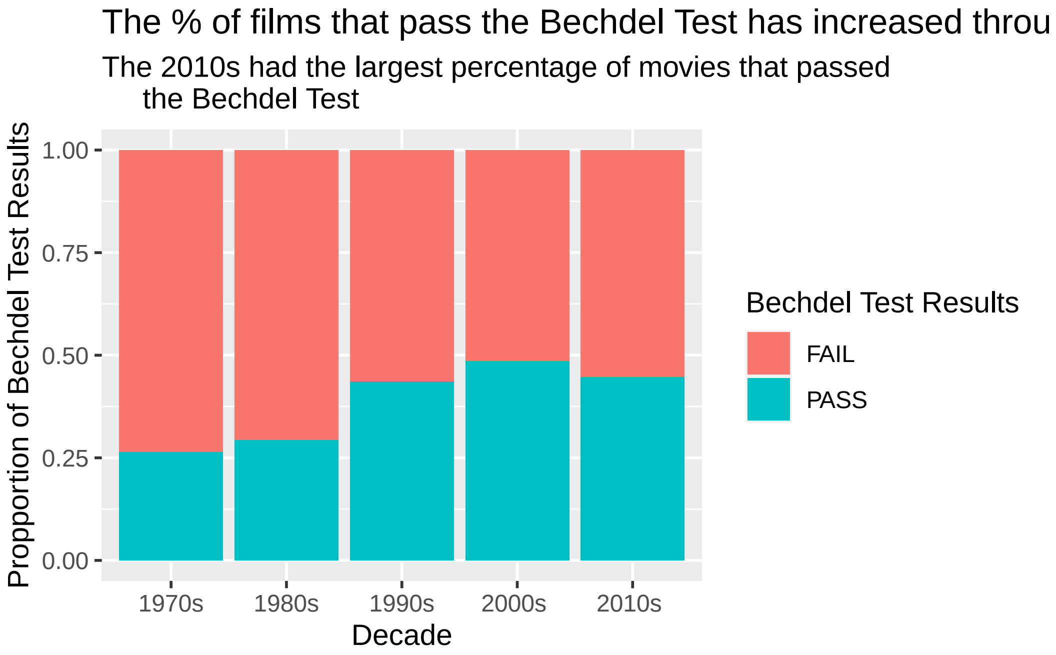
Data

To explore our reserch quetsion, we will be using a dataset used in the FiveThirtyEight story titled "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women". The dataset includes observations from 1794 films that were released between 1970-2013. The data set includes huge box office hit films including "Die Hard: With a Vengeance" and David Fincher's "Se7en." The data set includes information detailing the title of the film, the year it was released, its domestic gross, budget and international gross (both accounting for inflation and without). Furthermore, for our urposes, the data set also includes two important columns regarding whether the movie passed or failed the Bechdel test. The column "binary" specifically states whether teh movie passed or failed the bechdel test in a binary fashion. The column "clean_test" goes a bit more in detail, regarding how the film failed the bechdel test or if it was unclear whether the movie passed or not. The clean test variable has five levels: ok (Passed), no women (No women in the film), dubious (unclear result), no talk(women did not talk to each other), and men (women only talked about men.)

We created new variables for the purposes of our analysis. For starters, we created a decades variable that detailed which decade the given film was released in. The variable ended up having 5 different levels: 1970s, 1980s, 1990s, 2000s and 2010s. We also created a new variable titled passfail which is essentially the same as our binary variable but instead uses dummy values to illustrated whether the movies passed (passfail=1) pr failed (passdfail=0) the Bechdel test.

We excluded 18 observations from our final analysis. These observations had no values for their total domestic gross and total budget. These are variables of interest that important variables that we want to control for thus the movies that did not have the values for these variables were removed. Since we still had 1776 left in our data set, our statustical analyses will not be greatly affected since we are still working with a large data set.

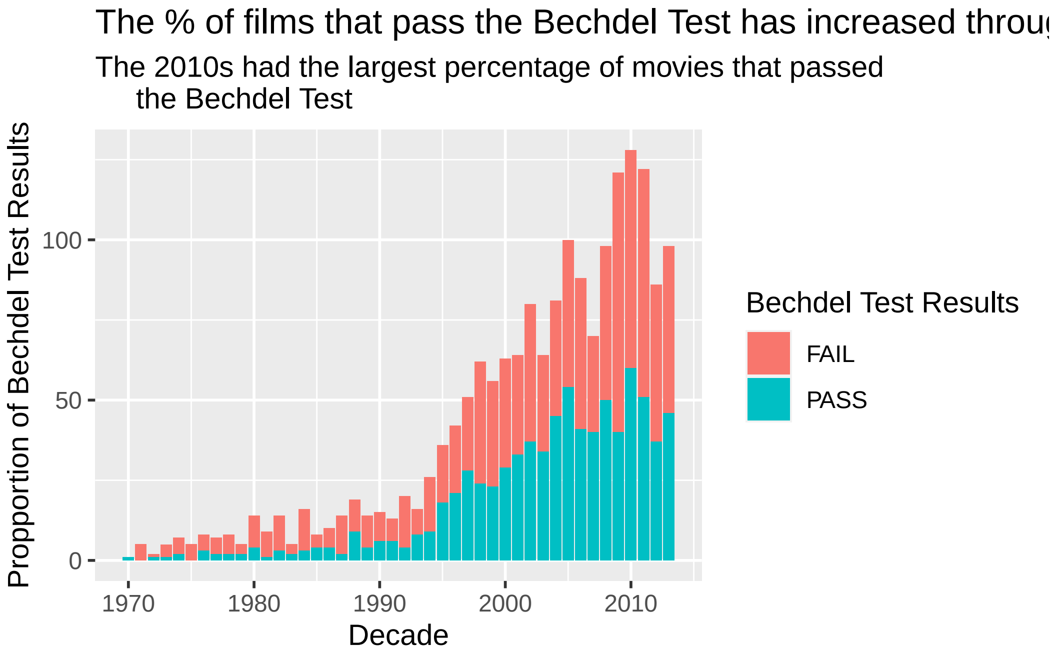
#Exploratory Data Analysis


The % of films that pass the Bechdel Test has increased throu

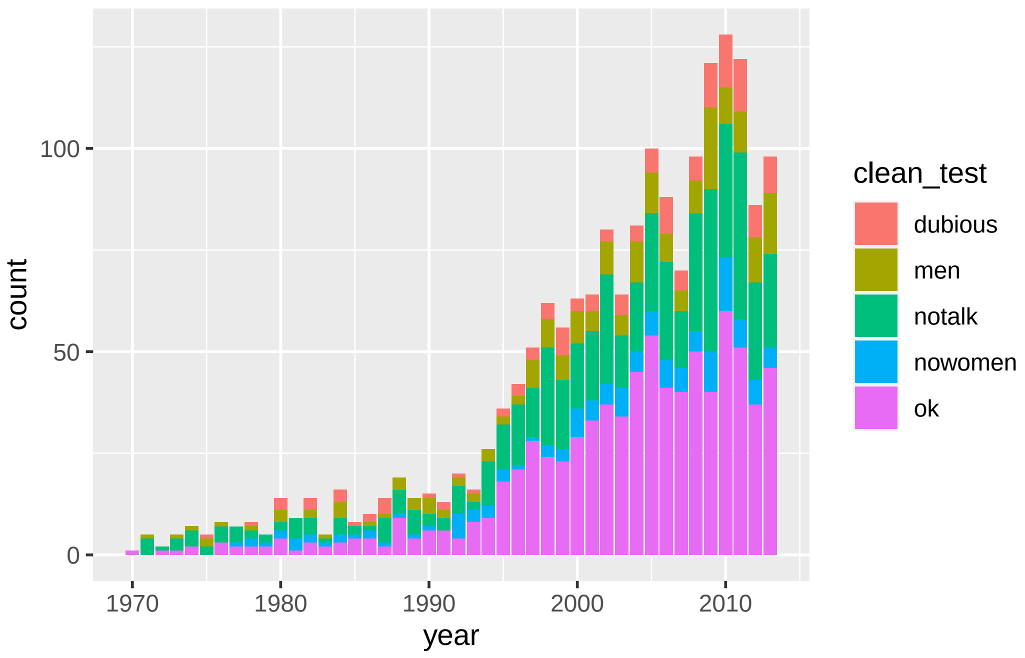
The 2010s had the largest percentage of movies that passed
the Bechdel Test

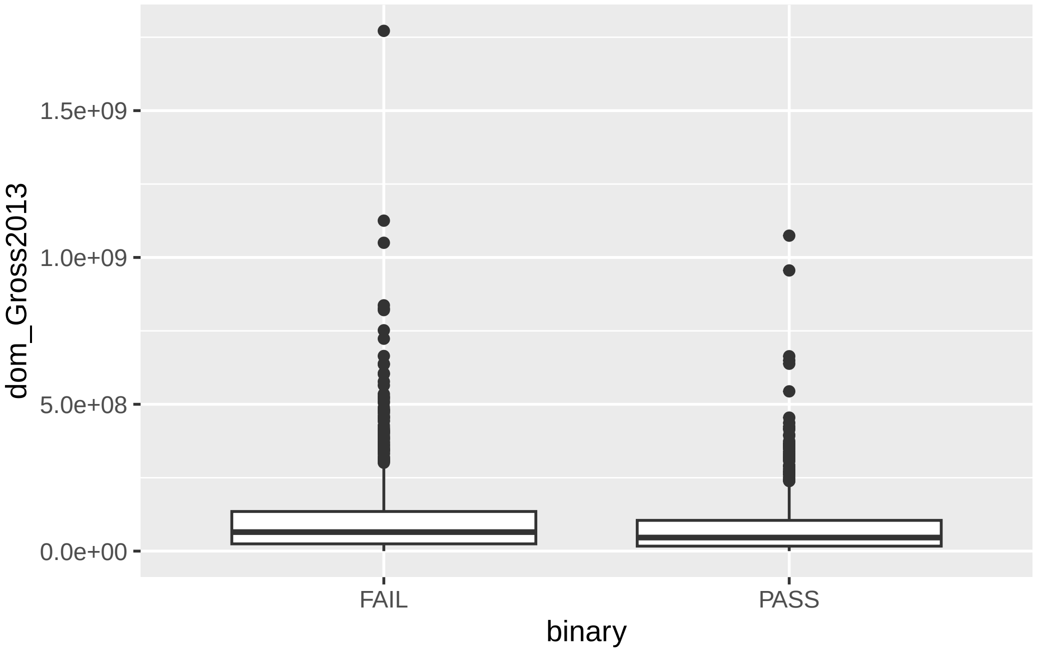
```
# A tibble: 5 x 2
  decade   number_of_movies
  <chr>               <int>
1 "1970s"                53
2 "1980s"               123
3 "1990s"               337
4 "2000s "              829
5 "2010s "              434
```

```
# A tibble: 5 x 2
  decade   passed_movies
  <chr>            <int>
1 "1970s"             14
2 "1980s"             36
3 "1990s"            147
4 "2000s "           403
5 "2010s "           194
```



We can see above that there is general upward trend across the decades with an increased percentage of movies that pass the Bechdel test. We see, specifically that the movies from our data set that premiered in the 2000s, about 48% of the films passed the Bechdel test. The 2010s performed in a similar fashion, with 45% of the films released passing the Bechdel test (it is important to note however that this data set only includes movies till 2013, thus it does not paint the full picture of female representation in film.) The decade with the lowest percentage of movies that passed the Bechdel test was teh 1970s, with only 25% passing.





#Methodolody

We are interested in running a regression model in order to evaluate whether the time period a given film was released.

First, we ran a logistic regression model to see if decade alone can predict whether a movie passes or fails the Bechdel test. We chose to run a logistic regression model and use the binary variable detailing whether the movie passed or did not pass the Bechdel tets. We chose to use the binary version detailing whether given dilm's performance on the Bechdel test rather than our clean_test variable. If we ran a logistic model with the clean_test as our outcome variable, we would have to use a multinomial regression model to test out our research question. A multinomial regression model would not made sense in this context the independence of irrelevant alternatives assumption would have been violated. This assumption assumes that, in a multinomial logistic regression model, the relative odds of choosing one option over another should not be influenced by the inclusion or exclusion of an additional option. This assumption does not apply in this case. This does not make sense since the inclusion or exclusion of a Bechdel test failing category could have an effect on our final analysis. For example, if a given film with plentiful female representation that was released in 2013 (which according to our hypothesis means it has a greater chance of passing the Bechdel test) was included our model but the only two categories taken accounted for whether "notalk" and "dubious," our model would predict it was it fit the dubious category. However, if the "ok" was included in teh mix, tsi woulc change our predictive probablity.

Furthermore, we also decided to use the year variable instead of our decade variable. We came to this conclusion by using the root mean squared error for our two possible models with. We only used the either year and decade for the purpose of this test.

Model 1:

$PassFail = \beta_0 + \beta_1(1980s)_i + \beta_2(1990s)_i + \beta_2(2000s) + \beta_3 I(2010s) + \epsilon_i$

Model 2: $PassFail = _0 + *1(year)+* $

We calculate the Root Mean Squared Errors for the two models to see which predictor variable would be the smartest to use:

We found that our first model had a RMSE for our first model was 0.8505231 and for our second model the RMSE was 0.8373541. Thus, we decided to use the second model for our analysis.

```

Call:
glm(formula = passfail ~ decade, family = "binomial", data = movies1_0)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1539  -1.0885  -0.8322   1.2011   1.6317

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

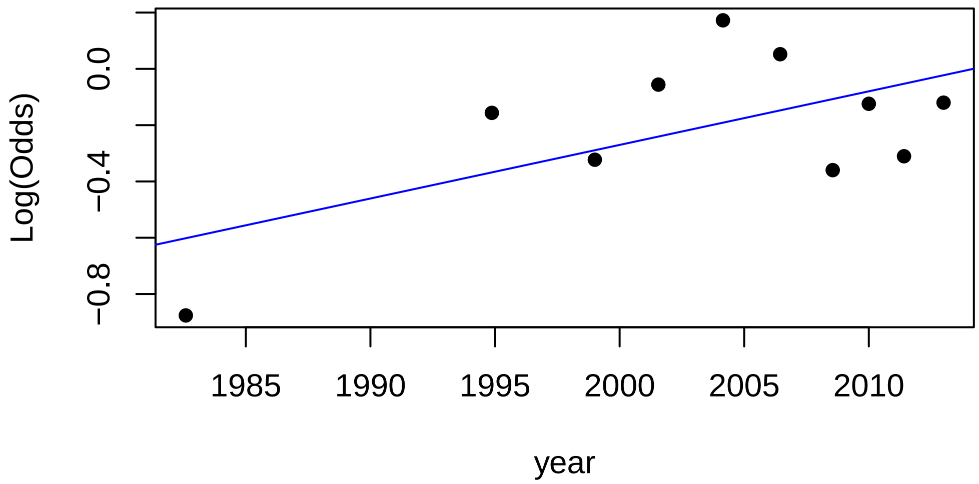
```
(Intercept)   -1.0245     0.3116  -3.288  0.00101 **
decade1980s    0.1421     0.3692   0.385  0.70033
decade1990s    0.7679     0.3304   2.324  0.02010 *
decade2000s    0.9690     0.3192   3.036  0.00240 **
decade2010s    0.8117     0.3262   2.489  0.01282 *
---
Signif. codes:  0 '***' 0.001 '**' 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: 2417.0  on 1771  degrees of freedom
AIC: 2427

Number of Fisher Scoring iterations: 4



Call:
glm(formula = passfail ~ year, family = "binomial", data = movies1_0)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.174  -1.116  -0.962   1.240   1.555

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -39.63178   10.97497  -3.611 0.000305 ***
year          0.01968    0.00548   3.592 0.000328 ***
---
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: 2428.9  on 1774  degrees of freedom
AIC: 2432.9

Number of Fisher Scoring iterations: 4


# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
```

```
1 rmse    standard       0.851
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       0.837
```

However, since we are using a logistic model, we are interested in seeing if it meets the losgistci regressio assumpttions. And since year is a continous variable we decided to see if it met our linearity condition.

```
  library(Stat2Data)
emplogitplot1(passfail ~ year,
              data = movies1_0,
              ngroups = 10)
```

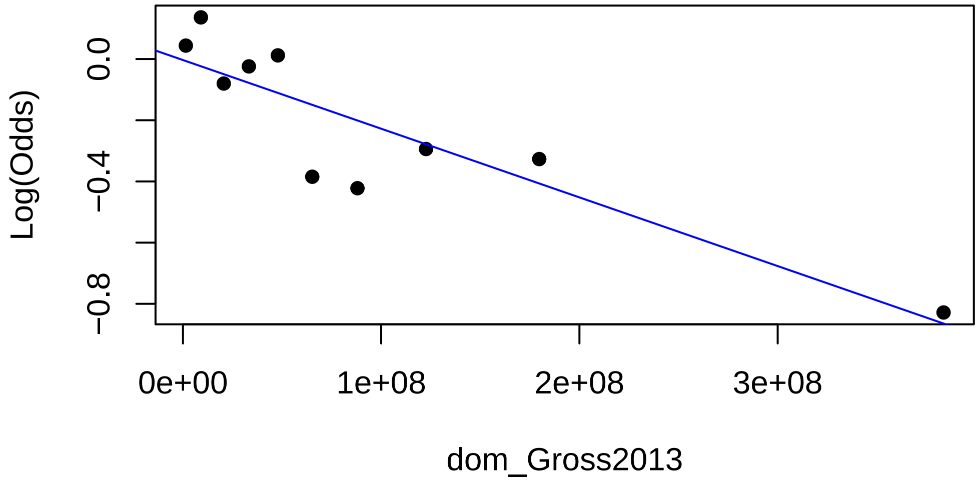


```
# nice
```

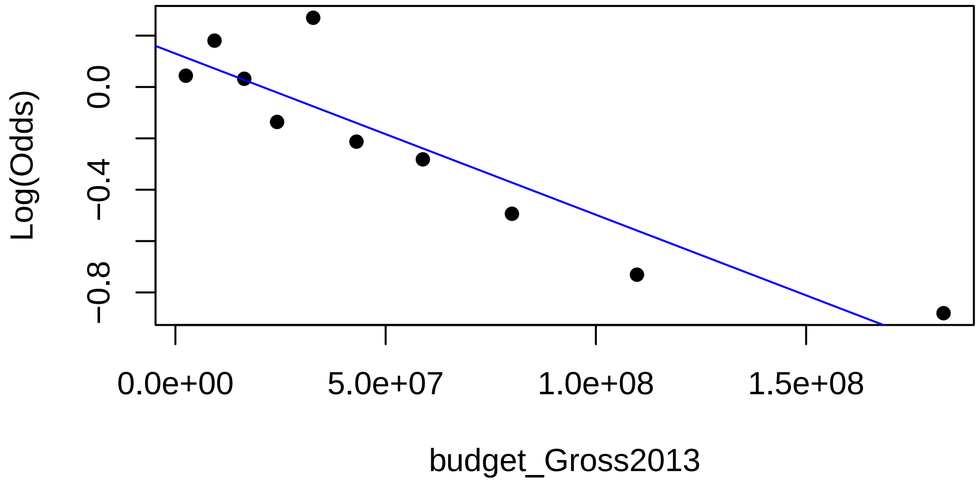
Points are not evenly scattered therefore we have decided that this is not an appropriate vraiable to use in our model and instead will use decade.

We also checked to see if the linearity condition was met for our other continous predictors.

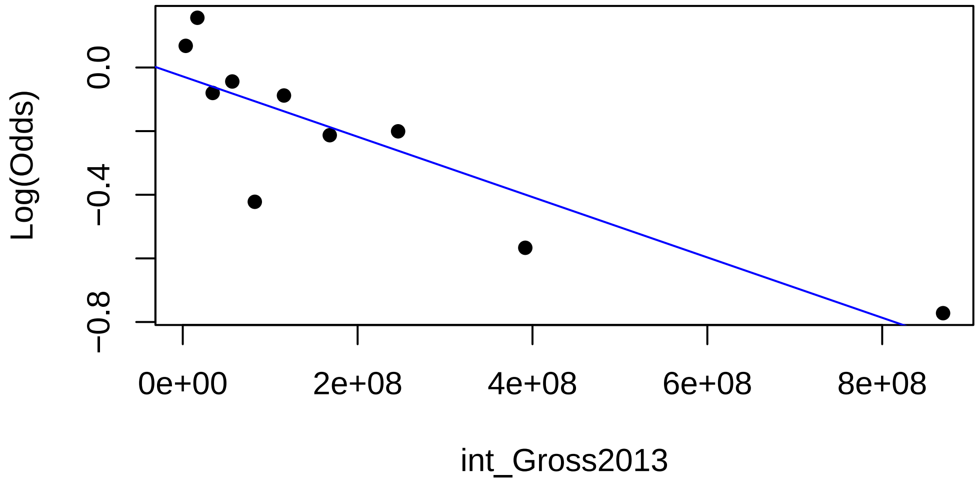
```
  library(Stat2Data)
emplogitplot1(passfail ~ dom_Gross2013,
              data = movies1_0,
              ngroups = 10)
```



```
  library(Stat2Data)
emplogitplot1(passfail ~ budget_Gross2013,
              data = movies1_0,
              ngroups = 10)
```



```
 library(Stat2Data)
emplogitplot1(passfail ~ int_Gross2013,
              data = movies1_0,
              ngroups = 10)
```



```
# nice
```

TALK ABOUT INDEPENDECE HERE:

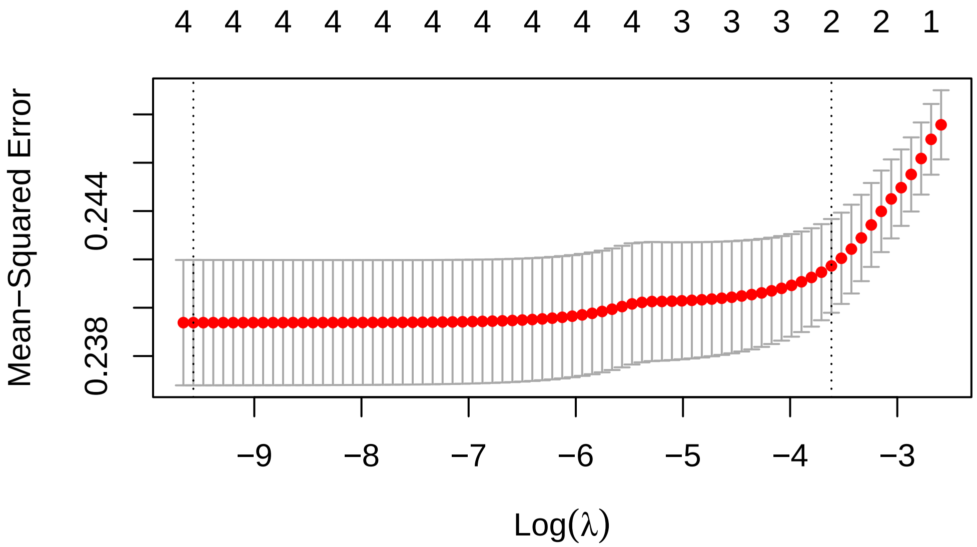
Furthermore, to further evaluate which variables to include in our data set, we ran a LASSO to determine if our variables of inteerst should be included in our model. They were all included thus we decided to include them in our model.

This is our proposed model:

$PassFail = _0 + _1(year)+ _2(budgetGross13) + _3(domGross13) + *3(intGross13) +*
$

```
[1] 6.988257e-05
```

```
5 x 1 sparse Matrix of class "dgCMatrix"
                             s0
(Intercept)       .
year              4.095117e-03
budget_Gross2013 -1.698329e-09
dom_Gross2013    -7.394474e-10
int_Gross2013     3.479047e-10
```



Results

First, we ran a logistic regression model to see if decade alone can predict whether a movie passes or fails the bechdel test.

$DomesticGross = \beta_0 + \beta_1(Budget)_i + \beta_2(Year)_i + \beta_2(Year) + \beta_3 I(BinaryResult) + \epsilon_i$

We will run a hypothesis at the $a = 0.05$ level.

Nulll hypothesis:

$H_0 : \beta_3 = 0$

There is not sufficient evidence to suggest that the year in which a movie premiered is associated with differential odds of the moving passing the Bechdel test, while controlling for all of the variables listed in the previous section.

Alternative Hypothesis:

$H_1 : \beta_3 \neq\ 0$

There is not sufficient evidence to suggest that the year in which a movie premiered is associated with differential odds of the moving passing the Bechdel test, while controlling for all of the variables listed in the previous section.

```
m2 <- glm(passfail ~ year + budget_Gross2013 +dom_Gross2013 + int_Gross2013,
          data = movies1_0,
          family = "binomial")
summary(m2)
```

```

Call:
glm(formula = passfail ~ year + budget_Gross2013 + dom_Gross2013 +
    int_Gross2013, family = "binomial", data = movies1_0)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3526  -1.1061  -0.8161   1.1743   1.9315

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.378e+01  1.245e+01  -2.714  0.00665 **
year              1.697e-02  6.211e-03   2.732  0.00630 **
budget_Gross2013 -7.472e-09  1.247e-09  -5.990  2.1e-09 ***
dom_Gross2013    -3.486e-09  1.281e-09  -2.720  0.00653 **
int_Gross2013     1.628e-09  5.842e-10   2.787  0.00533 **
---
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: 2374.5  on 1771  degrees of freedom
AIC: 2384.5

Number of Fisher Scoring iterations: 4
```

```
exp( 1.697e-02 )
```

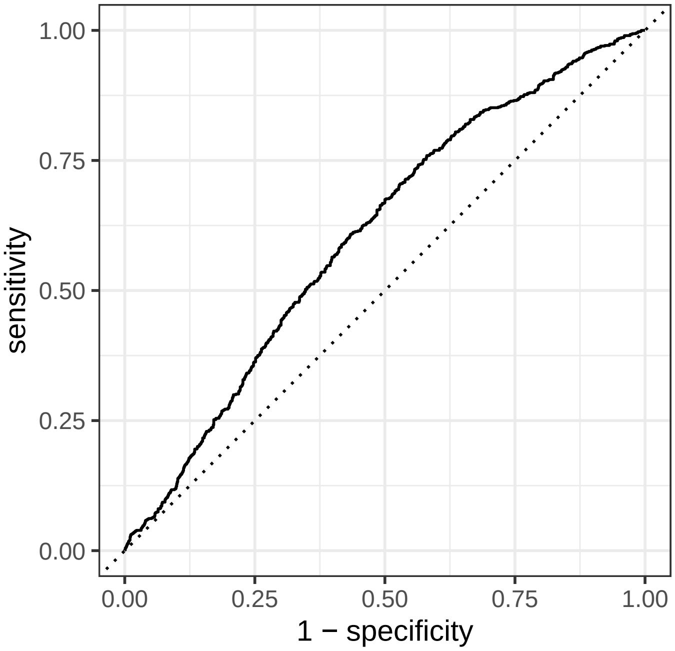
```
[1] 1.017115
```

We will be conducing a z test for this formal hypothesis test. The z statistic is 2.732. Z has a standard normal distribution under the null hypothesis in this test.

We reject the null hypothesis in this case since our p value is less than 0.05. While controlling for the variables in our model, for every additional increase in year a given film was released, the odds of the movie passing the Bechdel is predicted to be multiplied by 1.017115.

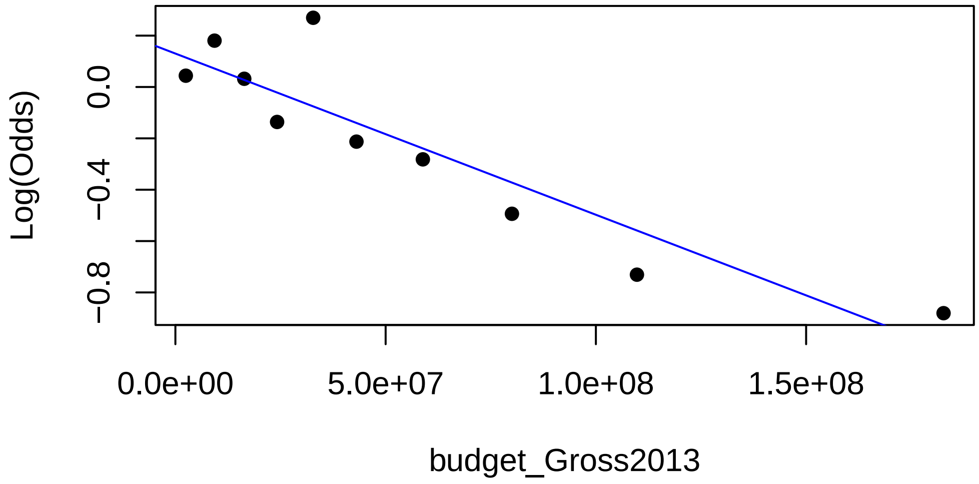
```
         0   1
  Fail 681 438
  Pass 301 356
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.607
```



The AUC (area under the curve) can be used to assess how well we are predicting, and summarizes the entire ROC curve. An AUC of 0.5 implies that the model is no better than a coin flip - an AUC of 1 implies a perfect fit.

```
  library(Stat2Data)
emplogitplot1(passfail ~ budget_Gross2013,
              data = movies1_0,
              ngroups = 10)
```



```
# nice
```

##Sources

https://socialsciences.ucla.edu/wp-content/uploads/2022/03/UCLA-Hollywood-Diversity-Report-2022-Film-3-24-2022.pdf

https://www.nytimes.com/2017/02/21/movies/women-protagonists-movies-2016.html
https://www.tandfonline.com/doi/pdf/10.1080/14680777.2016.1234239?needAccess=true

https://bechdeltest.com/

https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/