# STA 141A Final Project

Ian Xu, Nazil Luqman, Victoria Yeo, Jeffrey Ugochukwu

12/18/2020

# Contributions

Victoria Yeo: Data Analysis, Interpretation Ian Xu: Interpretation,Linear Regression Modeling, Project Manager Jeffrey Ugochukwu: Data Analysis, Conclusion Nazil Luqman: Data Management, Interpretation, Aesthetics

# Introduction

With anime expanding internationally over the last couple decades, one might think that the anime industry is fairing well. Yet, in "11+ Anime Comapnies and Studios that went Bankrupt in the Last 4 Decades," Theo Ellis, found of Anime Motivation, states that the "anime indsutry is a mess" due to reasons like high piracy of anime, low budgets, the average anime costing $1M dollars to produce, and some companies having difficulty paying employees and staying afloat. Ellis notes that "In the last 5+ years, anime companies have gone bankrupt more than ever."

The problem for anime studios is that they face a constraint one. With limited resources and hundreds of different production options, among which inlude 43 different genres and a virtually infinite number genre combinations, anime studios must determine what genre animes to produce that will garner enough interest to help them stay afloat and generate a positive rate of return. The objective of this project is decipher the data to help anime studios make better decisions in determining what animes to produce.

# Questions of Interest

There are two key questions that we hope to answer:

- What effect do each of the 43 different genres have on anime popularity?
- What model is best at predicting future observations of anime popularity?

# Dataset

Our dataset, found on Kaggle, is based on MyAnimeList.net, a website that allows members to score anime and manga. While the dataset includes many variables for each observation (6668 total observations), we found that some variables were inappropriate to include in this study (e.g. one variable contained website links to respective anime posters). Therefore, we have reduced the dataset to include the following variables:

- Popularity of Anime ($members$)
- Medium type ($type$) [Movie, Music, ONA, OVA, Special, TV]
- Anime source ($source$) [4-koma manga, Book, Card game, Digital manga, Game, Light novel, Manga, Music, Novel, Original, Other, Picture book, Radio, Visual novel, Web manga]
- Current airing status ($status$) [Currently Airing, Finished Airing]
- Age rating ($rating$) [G - All Ages, None, PG - Children, PG - 13 - Teens 13 or older, R - 17+ (violence & profanity), R+ - Mild Nudity, Rx - Hentai]
- Anime genre ($genre$) [Action, Adventure, Cars,Comedy, Dementia, Demons, Drama, Ecchi, Fantasy, Game, Harem, Hentai, Historical, Horror, Josei, Kids, Magic, Martial Arts, Mecha, Military, Music, Mystery, Parody, Police, Psychological, Romance, Samurai, School, Sci-Fi, Seinen, Shoujo, Shoujo Ai, Shounen, Shounen Ai, Slice of Life, Space, Sports, Super Power, Supernatural, Thriller, Vampire, Yaoi, Yuri, None]
- Year anime aired ($yearAired$) [1942, 1943, 1944, 1945, 1957, 1958,..., 2018]
- Number of episodes ($episodes$)
- Average rating out of 10 ($score$)
- Number of producers ($producerCount$)
- Number of licensors ($licensorCount$)
- Number of studios ($studioCount$)
- Number of genres anime encompasses ($genreCount$)

- Airing time, number of minutes ($duration$)

Because $type$, $source$, $status$, and $rating$ are categorical variables, we have to create a "baseline" case to avoid perfect collinearity. The baseline case is:

- $type$ = Movie
- $source$ = 4-koma manga
- $status$ = Currently Airing
- $rating$ = G - All Ages
- $genre$ = None
- $yearAired$ = 1942

# Causal Inference

To answer the first question of what effects different genres have on popularity ($members$), we will be conducting causal inference. With this in mind, the primary focus is whether or not our model violates any of the Gauss Markov assumptions required for OLS to be BLUE. Let's first start with defining our $x$ variables (for simplification purposes).

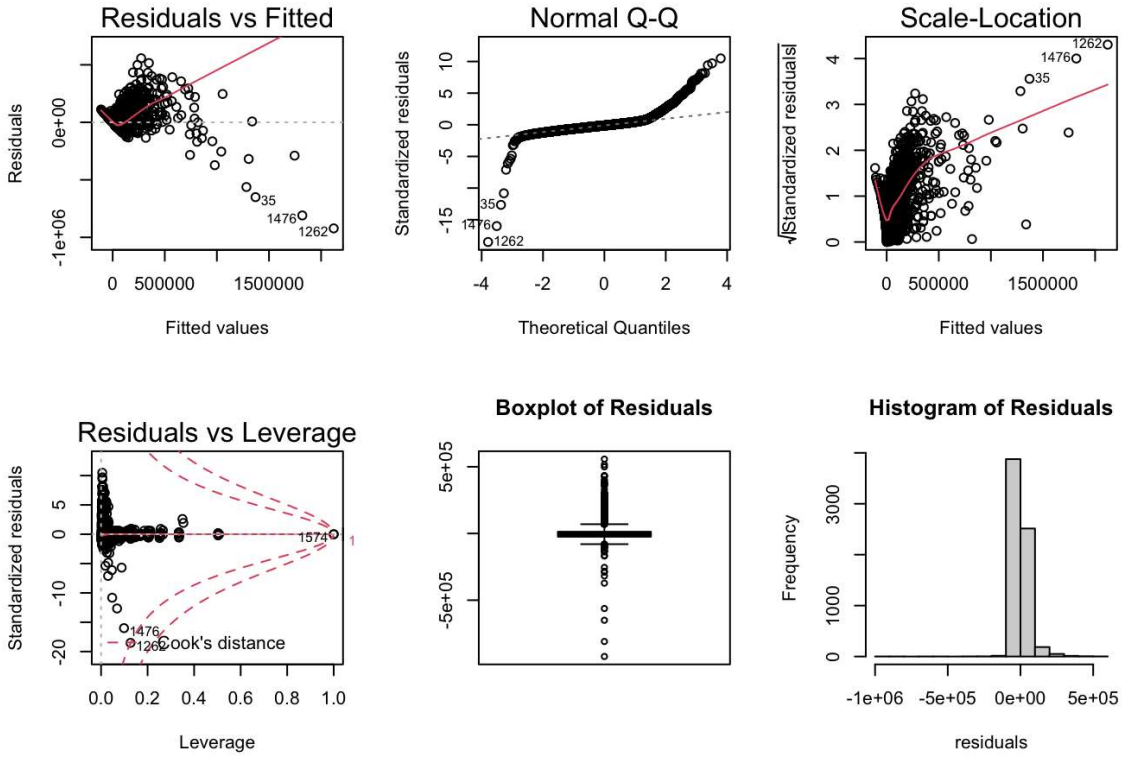
Let's let:

$$x_i = \begin{cases} \begin{cases} 1 \text{ for } i = 1, \dots, 5 & \text{if } type = \text{(Music, ONA, OVA, Special, TV), respectively} \\ 1 \text{ for } i = 6, \dots, 19 & \text{if } source = \text{(Book, Card game, Digital manga, Game, Light novel, Manga, Music, Novel, Original, Other, Picture book, Radio, Visual novel, Web manga), respectively} \\ 1 \text{ for } i = 20 & \text{if } status = \text{Finished Airing} \\ 1 \text{ for } i = 21, \dots, 26 & \text{if } rating = \text{(None, PG - Children, PG - 13 - Teens 13 or older, R - 17+ (violence \& profanity), R+ - Mild Nudity, Rx - Hentai), respectively} \\ 1 \text{ for } i = 27, \dots, 69 & \text{if } genre = \text{(Action, Adventure, Cars,Comedy, Dementia, Demons, Drama, Ecchi, Fantasy, Game, Harem, Hentai, Historical, Horror, Josei, Kids, Magic, Martial Arts, Mecha, Military, Music, Mystery, Parody, Police, Psychological, Romance, Samurai, School, Sci-Fi, Seinen, Shoujo, Shoujo Ai, Shounen, Shounen Ai, Slice of Life, Space, Sports, Super Power, Supernatural, Thriller, Vampire, Yaoi, Yuri), respectively} \\ 1 \text{ for } i = 70, \dots, 135 & \text{if } yearAired = \text{(1942, 1943, 1944, 1945, 1957, 1958,..., 2018), respectively} \end{cases} , 0 \text{ o.w.} \\ \begin{cases} episodes & \text{for } i = 136 \\ score & \text{for } i = 137 \\ producerCount & \text{for } i = 138 \\ licensorCount & \text{for } i = 139 \\ studioCount & \text{for } i = 140 \\ duration & \text{for } i = 141 \end{cases} \end{cases}$$

With that, our initial model is:

$$members = \alpha + \sum_{i=1}^{141} \beta_i x_i$$

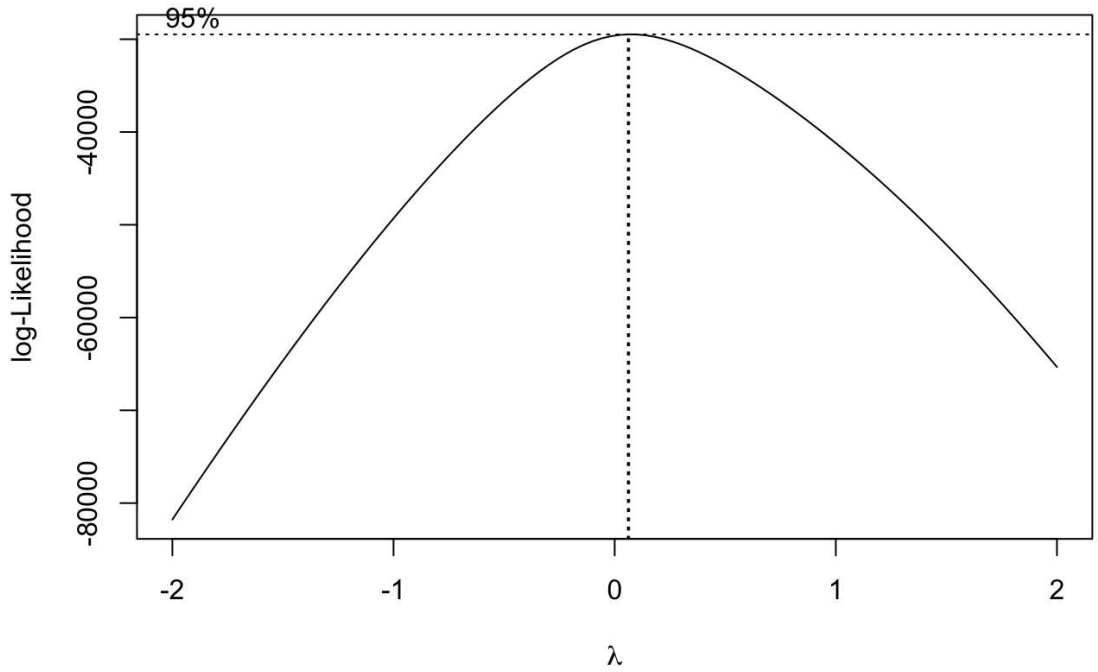
As stated above, our first focus is whether or not our model violates the Gauss Markov assumptions. If it does, OLS will not be BLUE and the estimated coefficients of genre would not represent the true causal effect of having an anime with some specific genre. If we do find violations, we can make necessary adjustments to our model like applying Box-Cox transformation or perhaps even using a different estimator (like weighted OLS).

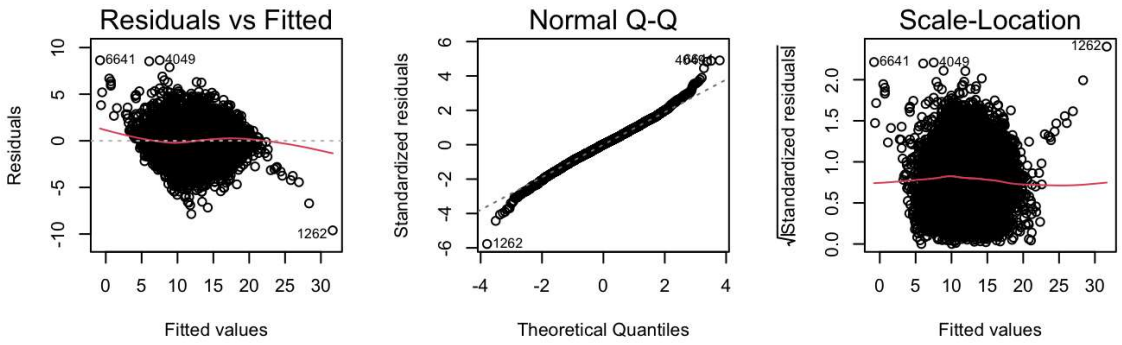


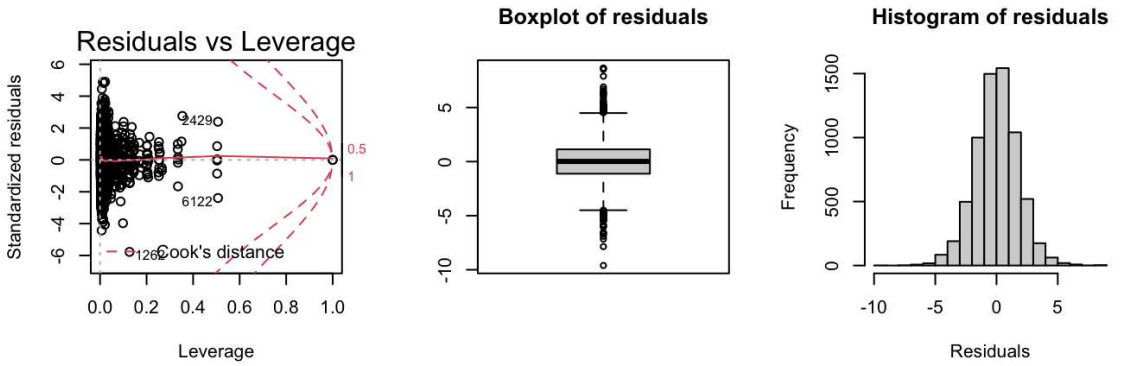

It is clear from the diagnostic plots that OLS is not BLUE in this case. It is clear from the Normal Q-Q plot that the residuals are not normally distributed. From the Histogram of Residuals and Boxplot of Residuals, the residuals seem to be skewed as well. The Residuals vs Fitted and Scale-Location graph show that the residuals suffer from heteroskedasticity as well. Furthermore, from the Residuals vs Leverage plot, it appears that there are some high leverage points which likely bias our results negatively.

# Applying Box-Cox Transformation

To address these concerns, we will apply a Box-Cox transformation. The best power ($\lambda$) is estimated by maximizing log-likelihood.

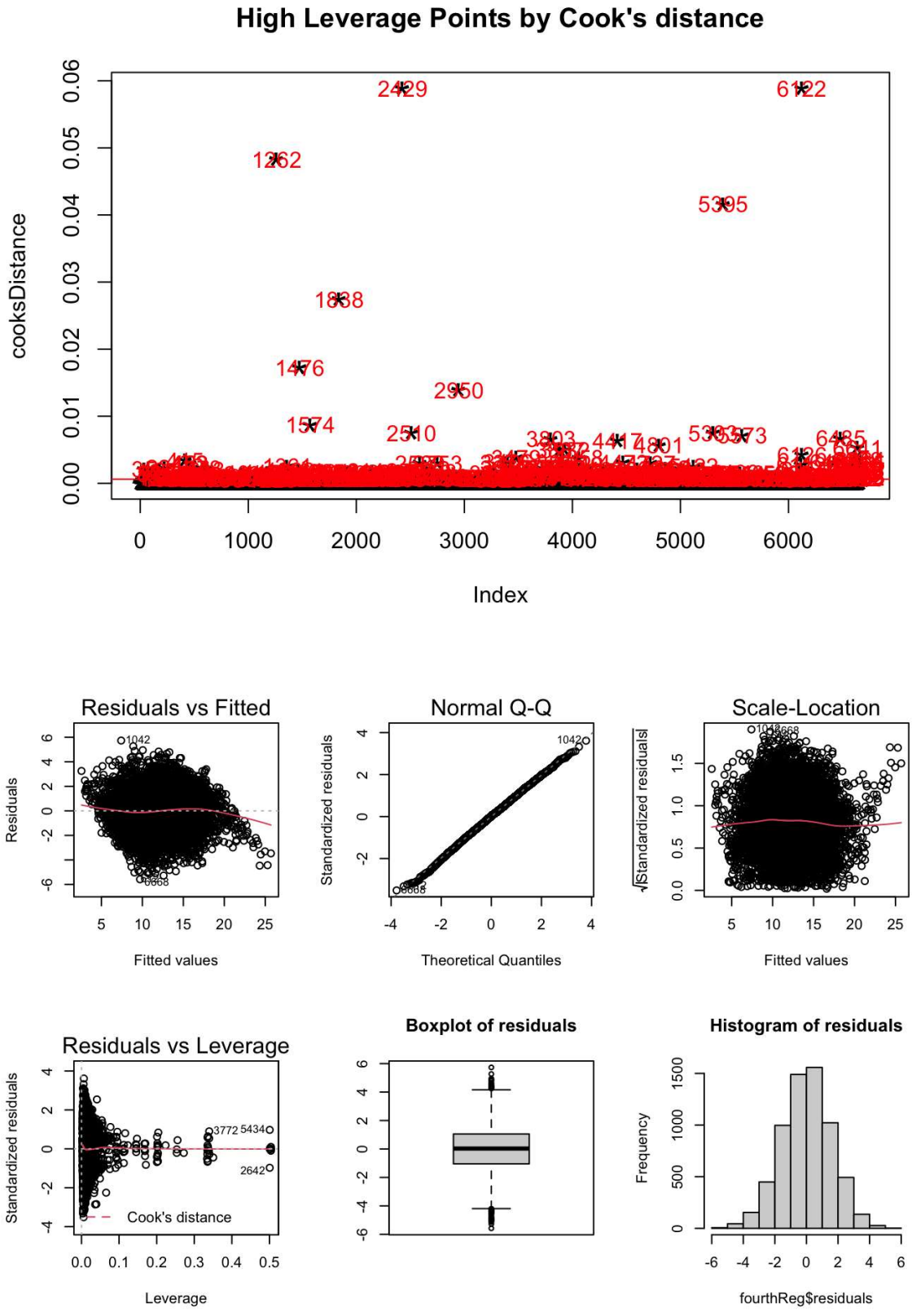



# Outlier removal

After applying the Box-Cox transformation, we will remove high leverage points, so that the estimated parameter for each genre better reflect the true value. The method we will use to remove high leverage points is measure the Cook's Distance of each observation. Then, using a standard 4/(n-k-1) criterion, we will remove all observations with a Cook's Distance greater than 6.1293x10^(-4). These observations are high leverage points.




After removing the high leverage points, the diagnostic plots appear much better. The model no longer appears to suffer from non-normally distributed errors, heteroskedasticity, and high leverage points. It is now appropriate to analyze the coefficients of each genre.

```
##
## Call:
## lm(formula = members_boxcox[-sort(highLeveragePoints)] ~ score +
##     source + status + favorites + producerCount + licensorCount +
##     studioCount + genreCount + Action + Adventure + Cars + Comedy +
##     Dementia + Demons + Drama + Ecchi + Fantasy + Game + Harem +
##     Hentai + Historical + Horror + Josei + Kids + Magic + `Martial Arts` +
##     Mecha + Military + Music + Mystery + Parody + Police + Psychological +
##     Romance + Samurai + School + `Sci-Fi` + Seinen + Shoujo +
##     `Shoujo Ai` + Shounen + `Shounen Ai` + `Slice of Life` +
##     Space + Sports + `Super Power` + Supernatural + Thriller +
##     Vampire + Yaoi + Yuri + duration + episodes + type + rating +
##     factor(yearAired), data = dataMinusOutliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5773 -1.0485  0.0293  1.0550  5.7327
##
## Coefficients: (43 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -2.377e+00  1.157e+00  -2.055 0.039900 *
## score                            1.637e+00  2.981e-02  54.924  < 2e-16 ***
## sourceBook                      -5.883e-02  3.245e-01  -0.181 0.856136
## sourceCard game                 -1.030e+00  2.742e-01  -3.757 0.000174 ***
## sourceDigital manga              2.385e+00  9.289e-01   2.567 0.010279 *
## sourceGame                       2.008e-01  1.437e-01   1.398 0.162237
## sourceLight novel                8.952e-01  1.384e-01   6.468 1.07e-10 ***
## sourceManga                      1.138e-01  1.236e-01   0.920 0.357490
## sourceMusic                     -2.177e-01  3.043e-01  -0.715 0.474386
## sourceNovel                     -7.241e-02  1.573e-01  -0.460 0.645362
## sourceOriginal                  -2.230e-01  1.250e-01  -1.784 0.074496 .
## sourceOther                      3.941e-02  1.761e-01   0.224 0.822876
## sourcePicture book              -2.221e+00  2.728e-01  -8.142 4.65e-16 ***
## sourceRadio                      2.029e-01  1.135e+00   0.179 0.858118
## sourceVisual novel               5.681e-01  1.481e-01   3.835 0.000127 ***
## sourceWeb manga                  3.090e-01  1.999e-01   1.546 0.122173
## statusFinished Airing            3.927e-01  1.451e-01   2.707 0.006811 **
## favorites                        1.669e-04  8.209e-06  20.332  < 2e-16 ***
## producerCount                    2.107e-01  1.207e-02  17.450  < 2e-16 ***
## licensorCount                    1.087e+00  3.842e-02  28.289  < 2e-16 ***
## studioCount                     -4.757e-02  6.881e-02  -0.691 0.489453
## genreCount                       1.658e-01  1.466e-02  11.313  < 2e-16 ***
## Action                                  NA         NA      NA       NA
## Adventure                               NA         NA      NA       NA
## Cars                                    NA         NA      NA       NA
## Comedy                                  NA         NA      NA       NA
## Dementia                                NA         NA      NA       NA
## Demons                                  NA         NA      NA       NA
## Drama                                   NA         NA      NA       NA
## Ecchi                                   NA         NA      NA       NA
## Fantasy                                 NA         NA      NA       NA
## Game                                    NA         NA      NA       NA
## Harem                                   NA         NA      NA       NA
## Hentai                                  NA         NA      NA       NA
## Historical                              NA         NA      NA       NA
## Horror                                  NA         NA      NA       NA
## Josei                                   NA         NA      NA       NA
## Kids                                    NA         NA      NA       NA
## Magic                                   NA         NA      NA       NA
## `Martial Arts`                          NA         NA      NA       NA
## Mecha                                   NA         NA      NA       NA
## Military                                NA         NA      NA       NA
```

```
## Music                                    NA         NA      NA       NA
## Mystery                                  NA         NA      NA       NA
## Parody                                   NA         NA      NA       NA
## Police                                   NA         NA      NA       NA
## Psychological                            NA         NA      NA       NA
## Romance                                  NA         NA      NA       NA
## Samurai                                  NA         NA      NA       NA
## School                                   NA         NA      NA       NA
## `Sci-Fi`                                 NA         NA      NA       NA
## Seinen                                   NA         NA      NA       NA
## Shoujo                                   NA         NA      NA       NA
## `Shoujo Ai`                              NA         NA      NA       NA
## Shounen                                  NA         NA      NA       NA
## `Shounen Ai`                             NA         NA      NA       NA
## `Slice of Life`                          NA         NA      NA       NA
## Space                                    NA         NA      NA       NA
## Sports                                   NA         NA      NA       NA
## `Super Power`                            NA         NA      NA       NA
## Supernatural                             NA         NA      NA       NA
## Thriller                                 NA         NA      NA       NA
## Vampire                                  NA         NA      NA       NA
## Yaoi                                     NA         NA      NA       NA
## Yuri                                     NA         NA      NA       NA
## duration                           -3.450e-03  1.267e-03  -2.724 0.006473 **
## episodes                           -6.714e-03  9.883e-04  -6.793 1.20e-11 ***
## typeMusic                          -5.909e-01  2.291e-01  -2.579 0.009932 **
## typeONA                            -5.757e-01  1.258e-01  -4.578 4.78e-06 ***
## typeOVA                             7.144e-04  9.576e-02   0.007 0.994048
## typeSpecial                        -5.451e-01  1.010e-01  -5.398 6.99e-08 ***
## typeTV                              7.735e-01  9.659e-02   8.008 1.37e-15 ***
## ratingNone                         -9.477e-02  1.938e-01  -0.489 0.624756
## ratingPG - Children                 4.776e-01  9.474e-02   5.041 4.75e-07 ***
## ratingPG-13 - Teens 13 or older     1.615e+00  6.681e-02  24.174  < 2e-16 ***
## ratingR - 17+ (violence & profanity)  2.067e+00  8.861e-02  23.329  < 2e-16 ***
## ratingR+ - Mild Nudity              2.265e+00  9.658e-02  23.455  < 2e-16 ***
## ratingRx - Hentai                   7.602e-01  1.361e-01   5.588 2.39e-08 ***
## factor(yearAired)1962               6.169e-01  1.951e+00   0.316 0.751891
## factor(yearAired)1963              -1.631e+00  1.455e+00  -1.121 0.262418
## factor(yearAired)1964              -8.991e-01  1.458e+00  -0.617 0.537412
## factor(yearAired)1966              -8.282e-01  1.595e+00  -0.519 0.603562
## factor(yearAired)1969              -2.176e+00  1.596e+00  -1.363 0.172841
## factor(yearAired)1970              -1.555e+00  1.457e+00  -1.067 0.285920
## factor(yearAired)1971              -1.541e+00  1.384e+00  -1.114 0.265395
## factor(yearAired)1972              -4.627e-01  1.459e+00  -0.317 0.751140
## factor(yearAired)1973              -2.027e+00  1.303e+00  -1.556 0.119743
## factor(yearAired)1974              -1.931e+00  1.281e+00  -1.507 0.131803
## factor(yearAired)1975              -6.851e-01  1.339e+00  -0.512 0.608838
## factor(yearAired)1976              -2.881e+00  1.336e+00  -2.157 0.031075 *
## factor(yearAired)1977              -2.800e+00  1.250e+00  -2.241 0.025063 *
## factor(yearAired)1978              -1.780e+00  1.337e+00  -1.332 0.183047
## factor(yearAired)1979              -1.375e+00  1.227e+00  -1.120 0.262549
## factor(yearAired)1980              -1.311e+00  1.203e+00  -1.090 0.275809
## factor(yearAired)1981              -1.498e+00  1.166e+00  -1.285 0.198946
## factor(yearAired)1982              -1.856e+00  1.185e+00  -1.566 0.117319
## factor(yearAired)1983              -1.515e+00  1.177e+00  -1.287 0.198243
## factor(yearAired)1984              -2.246e+00  1.160e+00  -1.936 0.052893 .
## factor(yearAired)1985              -1.578e+00  1.161e+00  -1.359 0.174062
## factor(yearAired)1986              -1.417e+00  1.153e+00  -1.229 0.219107
## factor(yearAired)1987              -1.631e+00  1.148e+00  -1.421 0.155472
## factor(yearAired)1988              -1.468e+00  1.154e+00  -1.272 0.203424
## factor(yearAired)1989              -1.787e+00  1.150e+00  -1.555 0.120103
## factor(yearAired)1990              -1.675e+00  1.151e+00  -1.455 0.145677
## factor(yearAired)1991              -1.899e+00  1.146e+00  -1.657 0.097616 .
```

```
## factor(yearAired)1992               -1.375e+00  1.149e+00  -1.197 0.231204
## factor(yearAired)1993               -1.572e+00  1.150e+00  -1.367 0.171626
## factor(yearAired)1994               -1.130e+00  1.146e+00  -0.986 0.324134
## factor(yearAired)1995               -1.241e+00  1.149e+00  -1.081 0.279923
## factor(yearAired)1996               -1.387e+00  1.144e+00  -1.212 0.225559
## factor(yearAired)1997               -1.394e+00  1.148e+00  -1.214 0.224773
## factor(yearAired)1998               -1.472e+00  1.142e+00  -1.288 0.197693
## factor(yearAired)1999               -1.030e+00  1.141e+00  -0.903 0.366748
## factor(yearAired)2000                1.571e-01  1.144e+00   0.137 0.890792
## factor(yearAired)2001               -4.400e-01  1.138e+00  -0.387 0.699044
## factor(yearAired)2002               -7.681e-02  1.138e+00  -0.068 0.946182
## factor(yearAired)2003               -2.550e-01  1.137e+00  -0.224 0.822598
## factor(yearAired)2004               -7.351e-02  1.136e+00  -0.065 0.948404
## factor(yearAired)2005                2.442e-01  1.136e+00   0.215 0.829743
## factor(yearAired)2006                8.660e-02  1.134e+00   0.076 0.939147
## factor(yearAired)2007                3.577e-01  1.134e+00   0.315 0.752460
## factor(yearAired)2008                7.109e-01  1.135e+00   0.627 0.530965
## factor(yearAired)2009                8.120e-01  1.134e+00   0.716 0.473795
## factor(yearAired)2010                8.022e-01  1.133e+00   0.708 0.479006
## factor(yearAired)2011                9.869e-01  1.133e+00   0.871 0.383667
## factor(yearAired)2012                1.071e+00  1.133e+00   0.945 0.344624
## factor(yearAired)2013                1.117e+00  1.133e+00   0.986 0.324060
## factor(yearAired)2014                9.121e-01  1.132e+00   0.806 0.420471
## factor(yearAired)2015                6.544e-01  1.132e+00   0.578 0.563092
## factor(yearAired)2016                4.894e-01  1.131e+00   0.433 0.665313
## factor(yearAired)2017                3.893e-01  1.131e+00   0.344 0.730827
## factor(yearAired)2018                5.579e-01  1.137e+00   0.491 0.623773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.591 on 6289 degrees of freedom
## Multiple R-squared:  0.8027, Adjusted R-squared:  0.7999
## F-statistic: 290.7 on 88 and 6289 DF,  p-value: < 2.2e-16
```

To determine which genres are significant, we will use a standard $\alpha = 0.05$. However, we will also apply a Bonferroni Correction, which results in comparing the p-values to $\frac{0.05}{43} = 0.00116279$ (we divided by 43 because there are 43 total genres). Based on this threshold, there are no significant genres. This indicates that genres do not play a significant role in determining the popularity ( $\backslash \text{members}$) of an anime. It is important to note that our model is not perfect. We had limited data, and we did the best we could with the data.

# Predicting Members

With causal inference, we transformed our initial model and removed outliers from our dataset with the goal of meeting all Gauss Markov assumptions such that OLS is BLUE. We didn't care about model complexity since our goal was to produce unbiased, consistent, and efficient estiamtes of the effect that different genres had on popularity. Ultimately, due to the weakness of cross secitonal data, we weren't able to meet all the requirements for OLS to be BLUE.

# "All models are wrong, but some are useful"

With prediction, our priorities shift. We're not so much worried about whether or not a model is "wrong," but instead whether or not it is useful in predicting $members$. To find the best variables that predict $members$, we first randomly divide our full dataset into training data and testing data—each composing 50% of the full dataset. Next, using the training dataset and starting with the empty model, we will perform bidirectional stepwise selection using AIC and BIC. The total possible varibales to include in our model is the same as our first model from causal inference, excluding genres. We opted not to include genres to simplify the overall model. Furthermore, we removed $yearAired$ from consideration since animes released in the future will never have $yearAired$ values of the past.

Starting with the empty model and the **training dataset**, bidirectional stepwise selection minimizing **AIC**, results in the model:

$$members \sim favorites + producerCount + score + source + licensorCount + type + rating \\ + genreCount + episodes + studioCount + duration$$

Starting with the empty model and the **testing dataset**, bidirectional stepwise selection minimizing **BIC**, results in the model:

$$members \sim favorites + producerCount + score + licensorCount + source + type + genreCount + episodes$$

The difference between the model that minimizes AIC and the model that minimizes BIC is the inclusion of $rating$, $studioCount$, and $duration$ in the AIC model. Now, we must determine whether the inclusion of these variables in the AIC model is "worth" the complexity that they bring. To determine this, we will compare $adj.\ R^2$ values, which take into consideration the number of parameters. The AIC model (1) has the higher $adj.\ R^2$ value than the BIC model (2), which suggests that the inclusion of $rating$, $studioCount$, and $duration$ is worth the additional complexity.

| | *Dependent variable:* | |
|---|---|---|
| | members | |
| | (1) | (2) |
| Observations | 3,334 | 3,334 |
| Adjusted $R^2$ | 0.725 | 0.721 |
| Akaike Inf. Crit. | 82,388.540 | 82,419.980 |
| Bayesian Inf. Crit. | 82,602.450 | 82,585.000 |

To further test whether the AIC model is better than the BIC model at prediction, we will use both models to predict $members$ with the **testing data**. The $adj.\ R^2$ results for the AIC model and BIC model are close, but the results confirm that the AIC model (1) is indeed better than the BIC model (2).

| | *Dependent variable:* | |
|---|---|---|
| | stepwiseSelectionAIC | stepwiseSelectionBIC |
| | (1) | (2) |
| Observations | 3,334 | 3,334 |
| Adjusted $R^2$ | 0.747 | 0.743 |
| Akaike Inf. Crit. | 81,924.470 | 81,966.860 |
| Bayesian Inf. Crit. | 82,138.380 | 82,131.880 |

# Conclusion

When discussing the relationship between genre and popularity, we looked at the p-values of the our initial model and applied a Bonferroni correction where we set a threshold for a standard alpha of our choosing and divided that by the 43 genres that exist, so that we can have a defined alpha as our threshold. For the first time we applied the correction, we used an original alpha of 0.05 and divided that by 43 to give us an alpha of 0.00116279. We then checked to see if the model had significant genres for this threshold, but unfortunately, there wasn't a genre that met this requirement since it was greater than our calculated alpha. This would mean that genre isn't a significant variable overall, which could mean that users are indifferent towards the anime's genre itself being an impact on its popularity.

To assess the quality of our model, we took out genre and ran a stepwise regression on this model. We created 2 models based on Akaike information criterion (AIC) and Bayes information criterion (BIC) using half our data as training data. We then assessed the other half of our data based on this model. The results we got showed our model based on AIC was better fitting of our data based adjusted r-squared values.

# Bibliography

https://www.liveabout.com/what-is-anime-144982 (https://www.liveabout.com/what-is-anime-144982)
https://en.wikipedia.org/wiki/History_of_anime (https://en.wikipedia.org/wiki/History_of_anime) https://animemotivation.com/anime-companies-bankrupt/ (https://animemotivation.com/anime-companies-bankrupt/)
http://www.css.cornell.edu/faculty/dgr2/_static/files/R_html/Transformations.html
(http://www.css.cornell.edu/faculty/dgr2/_static/files/R_html/Transformations.html) http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/
(http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/)

```
knitr::opts_chunk$set(echo = TRUE)
library(MASS)
library(stargazer)
knitr::opts_chunk$set(echo = FALSE, warning=FALSE)

#Getting Our Data
library(data.table)
data <- read.csv("/Users/Ian/Documents/STA 141A/FinalProjectData.csv")
genres = c("Action","Adventure", "Cars", "Comedy", "Dementia", "Demons", "Drama","Ecchi","Fantasy","Game", "Hare
m", "Hentai", "Historical", "Horror", "Josei", "Kids", "Magic",  "Martial Arts",  "Mecha",   "Military", "Music",
"Mystery" ,      "Parody" , "Police",  "Psychological", "Romance", "Samurai", "School",  "Sci-Fi",  "Seinen","Sho
ujo", "Shoujo Ai", "Shounen", "Shounen Ai", "Slice of Life", "Space" , "Sports", "Super Power", "Supernatural",
"Thriller", "Vampire", "Yaoi", "Yuri" )
data[genres] = 0
attach(data)
#Inputting Genre
data$Action = ifelse(grepl("Action", genre), 1, 0)
data$Adventure = ifelse(grepl("Adventure", genre), 1, 0)
data$Cars = ifelse(grepl("Cars", genre), 1, 0)
data$Comedy = ifelse(grepl("Comedy", genre), 1, 0)
data$Dementia = ifelse(grepl("Dementia", genre), 1, 0)
data$Demons = ifelse(grepl("Demons", genre), 1, 0)
data$Drama = ifelse(grepl("Drama", genre), 1, 0)
data$Ecchi = ifelse(grepl("Ecchi", genre), 1, 0)
data$Fantasy = ifelse(grepl("Fantasy", genre), 1, 0)
data$Game = ifelse(grepl("Game", genre), 1, 0)
data$Harem = ifelse(grepl("Harem", genre), 1, 0)
data$Hentai = ifelse(grepl("Hentai", genre), 1, 0)
data$Historical = ifelse(grepl("Historical", genre), 1, 0)
data$Horror = ifelse(grepl("Horror", genre), 1, 0)
data$Josei = ifelse(grepl("Josei", genre), 1, 0)
data$Kids = ifelse(grepl("Kids", genre), 1, 0)
data$Magic = ifelse(grepl("Magic", genre), 1, 0)
data$`Martial Arts` = ifelse(grepl("Martial Arts", genre), 1, 0)
data$Mecha = ifelse(grepl("Mecha", genre), 1, 0)
data$Military = ifelse(grepl("Military", genre), 1, 0)
data$Music = ifelse(grepl("Music", genre), 1, 0)
data$Mystery = ifelse(grepl("Mystery", genre), 1, 0)
data$Parody = ifelse(grepl("Parody", genre), 1, 0)
data$Police = ifelse(grepl("Police", genre), 1, 0)
data$Psychological = ifelse(grepl("Psychological", genre), 1, 0)
data$Romance = ifelse(grepl("Romance", genre), 1, 0)
data$Samurai = ifelse(grepl("Samurai", genre), 1, 0)
data$School = ifelse(grepl("School", genre), 1, 0)
data$`Sci-Fi` = ifelse(grepl("Sci-Fi", genre), 1, 0)
data$Seinen = ifelse(grepl("Seinen", genre), 1, 0)
data$Shoujo = ifelse(grepl("Shoujo", genre), 1, 0)
data$`Shoujo Ai` = ifelse(grepl("Shoujo Ai", genre), 1, 0)
data$Shounen = ifelse(grepl("Shounen", genre), 1, 0)
data$`Shounen Ai` = ifelse(grepl("Shounen Ai", genre), 1, 0)
data$`Slice of Life` = ifelse(grepl("Slice of Life", genre), 1, 0)
data$Space = ifelse(grepl("Space", genre), 1, 0)
data$Sports = ifelse(grepl("Sports", genre), 1, 0)
data$`Super Power` = ifelse(grepl("Super Power", genre), 1, 0)
data$Supernatural = ifelse(grepl("Supernatural", genre), 1, 0)
data$Thriller = ifelse(grepl("Thriller", genre), 1, 0)
data$Vampire = ifelse(grepl("Vampire", genre), 1, 0)
data$Yaoi = ifelse(grepl("Yaoi", genre), 1, 0)
data$Yuri = ifelse(grepl("Yuri", genre), 1, 0)
firstReg = lm(members~score+source+status+favorites+producerCount+licensorCount+studioCount+genreCount+ Action +
 Adventure + Cars  +  Comedy  +  Dementia  +  Demons  +  Drama  + Ecchi  + Fantasy  + Game  +  Harem  + Hentai  +
Historical  +  Horror  +  Josei  +  Kids  +  Magic  +   `Martial Arts`  +   Mecha  + Military  + Music  +    Myst
```

```
ery + Parody + Police + Psychological + Romance + Samurai + School + `Sci-Fi` + Seinen + Sh
oujo + `Shoujo Ai` + Shounen + `Shounen Ai` + `Slice of Life` + Space + Sports + `Super Power` + S
upernatural + Thriller + Vampire + Yaoi + Yuri + duration + episodes + type + rating + factor(yearAired),
data = data)

par(mfrow=c(2,3))
plot(firstReg)
boxplot(firstReg$residuals, main = "Boxplot of Residuals")
hist(firstReg$residuals, main = "Histogram of Residuals", xlab = "residuals")
secondReg = boxcox(firstReg)
power=secondReg$x[which.max(secondReg$y)]
BCTransform <- function(y, lambda=0) {
     if (lambda == 0L) { log(y) }
     else { (y^lambda - 1) / lambda }
  }
members_boxcox = BCTransform(data$members, power)
thirdReg <- lm(members_boxcox~score+source+status+favorites+producerCount+licensorCount+studioCount+genreCount+ A
ction + Adventure + Cars + Comedy + Dementia + Demons + Drama + Ecchi + Fantasy + Game + Harem + He
ntai + Historical + Horror + Josei + Kids + Magic + `Martial Arts` + Mecha + Military + Music
+ Mystery + Parody + Police + Psychological + Romance + Samurai + School + `Sci-Fi` + Se
inen + Shoujo + `Shoujo Ai` + Shounen + `Shounen Ai` + `Slice of Life` + Space + Sports + `Super Po
wer` + Supernatural + Thriller + Vampire + Yaoi + Yuri + duration + episodes + type + rating + factor(y
earAired), data = data)

par(mfrow=c(2,3))
plot(thirdReg)
boxplot(thirdReg$residuals, main = "Boxplot of residuals")
hist(thirdReg$residuals, main = "Histogram of residuals", xlab = "Residuals")
cooksDistance <- cooks.distance(thirdReg)

# Plot the Cook's Distance using the traditional 4/n criterion
plot(cooksDistance, pch="*", cex=2, main="High Leverage Points by Cook's distance")  # plot cook's distance
abline(h = 4/(6668-141-1), col="red")  # add cutoff line
text(x=1:length(cooksDistance)+1, y=cooksDistance, labels=ifelse(cooksDistance>4/(6668-141-1), names(cooksDistanc
e),""), col="red")  # add labels
highLeveragePoints <- as.numeric(names(cooksDistance)[cooksDistance > (4/(6668-141-1))])
highLeveragePoints = c(highLeveragePoints,1193, 1360, 1504, 5918, 6135, 1142, 1300, 1997 )
dataMinusOutliers <- data[-sort(highLeveragePoints), ]

fourthReg <- lm(members_boxcox[-sort(highLeveragePoints)]~score+source+status+favorites+producerCount+licensorCou
nt+studioCount+genreCount+ Action + Adventure + Cars + Comedy + Dementia + Demons + Drama + Ecchi + Fan
tasy + Game + Harem + Hentai + Historical + Horror + Josei + Kids + Magic + `Martial Arts` +
Mecha + Military + Music + Mystery + Parody + Police + Psychological + Romance + Samurai +
School + `Sci-Fi` + Seinen + Shoujo + `Shoujo Ai` + Shounen + `Shounen Ai` + `Slice of Life` + Sp
ace + Sports + `Super Power` + Supernatural + Thriller + Vampire + Yaoi + Yuri + duration + episod
es + type + rating + factor(yearAired), data = dataMinusOutliers)

par(mfrow=c(2,3))
plot(fourthReg)
boxplot(fourthReg$residuals, main = "Boxplot of residuals")
hist(fourthReg$residuals, main = "Histogram of residuals")
summary(fourthReg)
set.seed(2020141)
trainIndex <- sample(6668, as.integer(6668*0.5))
trainingData= data[trainIndex,]
testingData = data[-trainIndex,]

stepwiseSelectionAIC <- step(lm(members~1, data = trainingData),
                         scope = ~score+source+status+favorites+producerCount+licensorCount+studioCount+genreCou
nt+ duration + episodes + type + rating,
                         direction = "both",  k = 2,
                         trace = 0)
```

```
stepwiseSelectionBIC <- step(lm(members~1, data = trainingData),
                          scope = ~score+source+status+favorites+producerCount+licensorCount+studioCount+genreCou
nt+ duration + episodes + type + rating,
                          direction = "both",  k = log(as.integer(6668*0.5)),
                          trace = 0)
stepwiseSelectionAIC$AIC <- AIC(stepwiseSelectionAIC)
stepwiseSelectionAIC$BIC <- BIC(stepwiseSelectionAIC)

stepwiseSelectionBIC$AIC <- AIC(stepwiseSelectionBIC)
stepwiseSelectionBIC$BIC <- BIC(stepwiseSelectionBIC)

stargazer(stepwiseSelectionAIC, stepwiseSelectionBIC, type = "html", keep.stat = c("aic", "bic","adj.rsq","n"), o
mit.table.layout = "tn")
testDataAIC <- lm(stepwiseSelectionAIC$model, data = testingData)
testDataBIC <-lm(stepwiseSelectionBIC$model, data = testingData)

testDataAIC$AIC <- AIC(testDataAIC)
testDataAIC$BIC <- BIC(testDataAIC)

testDataBIC$AIC <- AIC(testDataBIC)
testDataBIC$BIC <- BIC(testDataBIC)

stargazer(testDataAIC, testDataBIC, type = "html", keep.stat = c("aic", "bic","adj.rsq","n"), omit.table.layout =
"tn")
```