Russell Feinstein
Term Project

As a fan of movies, it is always interesting to see how movie ratings go, especially when movies are more or less in the public's eye. I want to determine if movie popularity and the number of votes a movie got on IMDB had an affect on its vote average. Thus, I will be running multiple linear regression to determine if either (or both) movie popularity and the number of votes had an effect on the vote average, after accounting for the other.

The data set is information taken from TMDB's open API and put together for download on Kaggle. It can be found here: https://www.kaggle.com/datasets/ashvanths/toprated-tmdb-movies. There are three relevant variables I will be using in this analysis:

- Response Variable: Vote Average – the average vote rating a movie achieved
- Explanatory Variables:
  - Movie Popularity – this is an amalgamated score composed of a variety of metrics in order to determine how 'popular' a movie is. Without going into much detail, it is based on a number of things: unique views on the website, number of ratings, number of favorites, and number of watched list additions.
  - Vote Count – The number of votes a movie received.

I plan to use Multiple Linear Regression here to determine what effect movie popularity and vote count had on, if any, on the vote average of a movie. Multiple linear regression is used when we are interested in the relationship between each explanatory variable and the dependent variable after accounting for the remaining explanatory variables. This allows us to quantify the relationship between our response variable and our explanatory variables. In addition, it provides us a tool for predicting the response of a new observation for a given set of x values.

We will first use an F test to determine if the overall model has significance. Then we will use individual t-tests to determine which explanatory variable the significance is attributed to.

First, lets formally test whether popularity and vote count together are significant predictors of vote average.

1. Set up the Hypotheses and select the alpha level:
   $H_0$: $\beta_{\text{popularity}} = \beta_{\text{vote count}} = 0$ (popularity and vote count are not significant predictors of vote average)
   $H_1$: $\beta_{\text{popularity}} \neq 0$ and/or $\beta_{\text{vote count}} \neq 0$ (at least one of the slope coefficients is different than 0; popularity and/or vote count are significant predictors/is a significant predictor of vote average)
   $\alpha = 0.05$

2. Select the appropriate test statistic:
   $F$ = MS Regression/MS Residual, df = 2, n-k-1 (997)

3. State the decision rule
   Determine the appropriate value for the F statistic using R.
   $F_{2,\,997,\,0.05} = 3.0048$
   Decision Rule: Reject $H_0$ if $F \geq 3.0048$
   Otherwise, do not reject $H_0$

4. Compute the test statistic
   Using R, we find the F-statistic to be 35.24
5. Conclusion
   Reject $H_0$ since 35.24 ≥ 3.0048. We have significant evidence at the $\alpha = 0.05$ level that popularity and vote count, when taken together, are significant predictors of vote average. That is, there is evidence of a linear association between vote count and popularity and vote average.

Now that we have determined that the overall model is significant, we must look to see which explanatory variable the significance could be attributed to. We will do this using a t-test for each explanatory variable while controlling for the remaining explanatory variables.

First, we will test popularity:

1. Set up the hypotheses and select the alpha level
   $H_0$: $\beta_{\text{popularity}} = 0$ (after controlling for vote count)
   $H_1$: $\beta_{\text{popularity}} \neq 0$ (after controlling for vote count)
   $\alpha = 0.05$

2. Select the appropriate test statistic:

$$t = \frac{\hat{\beta}_{\text{age}}}{SE_{\hat{\beta}_{\text{age}}}} \ \text{df} = n - k - 1$$

   where df = 997

3. State the decision rule:
   Determine the appropriate value for the t-statistic using R.
   $t_{.997,\ .025} = 1.9623$
   Decision Rule: Reject $H_0$ if $|t| \geq 1.9623$
   Otherwise, do not reject $H_0$

4. Compute the test statistic using R
   $t = 2.436$

5. Conclusion
   Reject $H_0$ since 2.436 ≥ 1.9623. We have significant evidence at the $\alpha = 0.05$ level that $\beta_{\text{popularity}} \neq 0$ after controlling for vote count. That is, popularity is a significant predictor of vote average after adjusting for vote count.

Then, we must do the same for vote count

1. Set up the hypotheses and select the alpha level
   $H_0$: $\beta_{\text{vote count}} = 0$ (after controlling for vote count)
   $H_1$: $\beta_{\text{vote count}} \neq 0$ (after controlling for vote count)
   $\alpha = 0.05$

2. Select the appropriate test statistic:

   $$t = \frac{\hat{\beta}_{\text{age}}}{SE_{\hat{\beta}_{\text{age}}}} \text{ df} = n - k - 1$$

   where df = 997

3. State the decision rule:
   Determine the appropriate value for the t-statistic using R.
   $t_{.997,\ .025}$ = 1.9623
   Decision Rule: Reject $H_0$ if $|t| \geq 1.9623$
   Otherwise, do not reject $H_0$

4. Compute the test statistic using R
   $t$ = 6.845

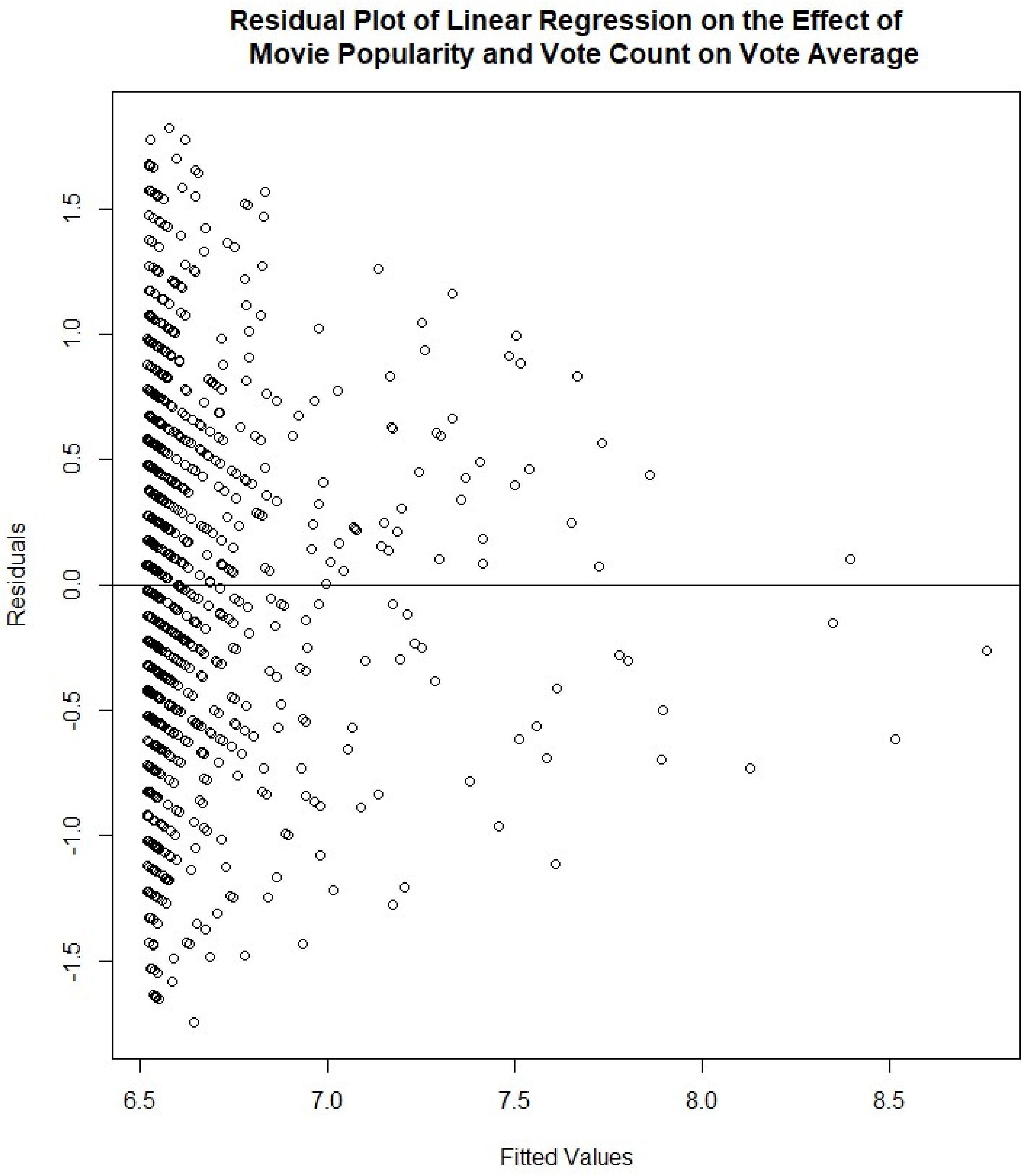
5. Conclusion
   Reject $H_0$ since 6.845 ≥ 1.9623. We have significant evidence at the $\alpha = 0.05$ level that $\beta_{\text{vote count}} \neq 0$ after controlling for popularity. That is, vote count is a significant predictor of vote average after adjusting for popularity.

In addition, it’s important we check for collinearity. The correlation coefficient between the two explanatory variables is only .3173 – this means we can safely assume the explanatory variables are not highly correlated and are not causing an issue in our multiple linear regression model.

The conclusion of this analysis is that both popularity and vote count are associated with the vote average, when accounting for the other. That is to say, when a movie is more popular and when a movie has more people voting on it, it tends to have a higher average vote.

A concern for me when running this is that it was done with a random sample of 1000 observations. A different random sampling could potentially hold different outcomes, which happened as I ran this experiment multiple times. I have also included in my R code the same multiple linear regression without sampling. Both the global F test and the individual t tests still hold true.

As for the assumptions, it is possible the constant variance does not hold. A residual plot is shown below. There is a clear decrease in the variance of the residuals as the fitted value increases.


Residual Plot of Linear Regression on the Effect of
Movie Popularity and Vote Count on Vote Average
Residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
6.5
7.0
7.5
8.0
8.5
Fitted Values