An Analysis of “Survivor” Reality TV Show Viewership

Anthony Rautmann

Introduction

"Survivor" is the title of a reality TV show that first aired in 2000. A group of "castaways" are transported to a remote location in a country, where they are separated into two tribes. The two tribes compete against each other in various challenges, and the losing tribe must vote off a team member. The last person standing earns the title Sole Survivor and wins a million dollars. Each week a different episode comes out with a different challenge, and all the while the castaways are living together in their tribes and are being filmed. Like with any reality TV show, the audience has a fun time discerning which events of the TV show are scripted and dramatized versus which events are true occurrences. I found myself a part of this audience in my former years watching "Survivor." A girl that attended the same high school as me was one of the castaways in the show, and she performed very well. I knew Survivor was a popular show in my hometown, but I didn't really have an idea of how it was received nationally, or globally, even. When I found a dataset reporting on "Survivor" viewers, my interest was reignited, and I was ready to perform data analysis to gain more insight about the TV show's viewership history. Some broad initial questions that jumped to my mind were, "*Did others find this show as interesting as I did*?" and, "*How many people other people have seen this show outside of my hometown*?"

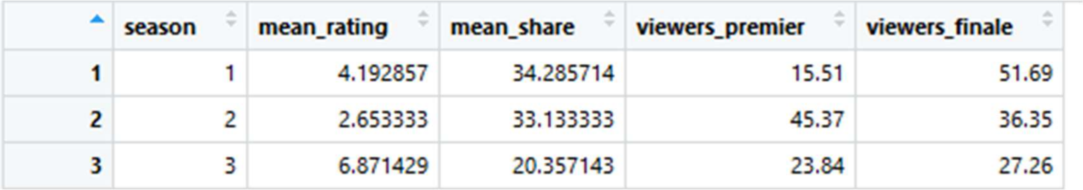
As I looked over the data set, I noticed that number of viewers for each of the premier and finale episodes were recorded. Also, the rating and share of each episode was recorded (more about this in the data section). I was particularly intrigued by the rating column, thinking about how it was recorded and who it was rated by. I thought, "*Is this rating system in the data* *source a good method for rating the show, and should I trust it*?" If it is a good rating system, then the rating scores of the episodes should have a significant effect on the number of viewers for the finale episode. I also wanted to see if the number of season premier viewers had a significant effect on the number of viewers for the finale episode, which would mean that as the number of viewers for the premier increases, the number of viewers for the finale would also increase. If there is a significant effect, it would indicate that people like the show because "Survivor" was able to retain viewers or add more viewers along the filming of the consecutive seasons.

Data

The data source I am using is posted on "gradientdescneding.com /survivor-data-from-the-tv-series-in-r/".[1] Data cleaning and data transformation was performed on two different csv files from the website, being "summary" and "viewers." The schema for these tables includes 28 variables of which only 4 I was interested in doing data analysis with. The others are rather meaningless categorical variables, locations, or dates. I did not include location in my analysis as a categorical variable because each season is filmed in a remote country, and each chosen country is interesting so that would not have an impact on viewership. The summary file consists of data recorded by season, and the viewers file consists of data recorded by episode. I did a mean function of the rating and share columns and grouped by season to obtain season data in the viewers file. Some episodes did not have records for rating or share in the viewers

[1] Rfordatascience. (n.d.). *Tidytuesday/readme.md at master · rfordatascience/tidytuesday*. GitHub. Retrieved December 15, 2021, from

file, so the row level entries here needed to be dropped in order for aggregation for each of the 40 seasons to have a mean rating and mean share row. After that, I joined the summary and viewers file to obtain a final data frame with the variables season, mean rating, mean share, premier viewers, and finale viewers (see *Figure 1*).



	season	mean_rating	mean_share	viewers_premier	viewers_finale
1	1	4.192857	34.285714	15.51	51.69
2	2	2.653333	33.133333	45.37	36.35
3	3	6.871429	20.357143	23.84	27.26

Figure 1. First three rows of the final data frame I used for analysis, generated in R

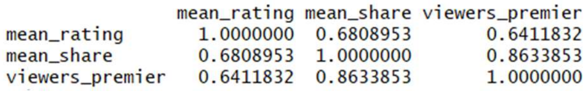
Season is the corresponding season of “Survivor” that was filmed. Mean rating is the average of the rating_18_49 column in the viewers file. A rating was recorded for each episode (a continuous numerical variable from 1-10) by a person in the age range of 18 to 49. Mean share is the average of the share_18_49 column in the viewers file. A share is the number of viewers aged 18 to 49 who were watching survivor during the air of an episode divided by the number of viewers who were watching TV in general during the air of an episode. Viewers_premier is the number of viewers, in millions, who were watching the given season premiere. Viewers_finale is the number of viewers, in millions, who were watching the given season finale. There are 40 rows total in the final data frame, one for each season of “Survivor.” Looking at this smaller sized data set, there does not seem to be any problems with any of the variables having outliers; however, there does seem to be a negative trend in all the variables as the seasons progress, especially in the last 10 seasons. Potentially, this could cause some multicollinearity issues between predictors if a multiple regression model is fit.

Analysis

I was interested in the explanatory power that mean share, mean rating, and premier viewers have on finale viewers. This question can be explored by fitting a Multiple Linear Regression Model (MLR) with three predictors and having finale viewers as the response variables. After fitting the model, I did a hypothesis test to ensure that at least one of the beta coefficients in the model is different from 0. The null hypothesis was that each beta was equal to 0, and the alternative hypothesis was that at least one beta was different from 0. To test, I constructed an ANOVA (analysis of variance) table comparing the full MLR with the null model (which is the model with no predictors and only and intercept estimate). The test produced a high F statistic and low P value, so I had sufficient evidence to reject the null hypothesis and conclude that at least one beta coefficient estimate was different from 0.

Entertaining the MLR model further, I wanted to investigate the individual impacts each predictor had on the response. This required me to test whether or not each predictor significantly impacted the model fit when all the other predictors are also included in the model. Three separate hypotheses tests with the null hypothesis being the given beta coefficient was equal to 0 and the alternative hypothesis being the given beta coefficient was not equal to 0 had to be conducted. To do this, I constructed the ANOVA table comparing the full MLR model to another MLR model without the testing beta in it. Each test resulted in a p-value that was significant, so I had sufficient evidence to reject all the null hypotheses in the three tests and conclude that each beta coefficient was significantly different from 0; therefore, the predictors mean rating, mean score, and premier viewers all significantly contributed to the full MLR model. The R-squared value for the full MLR model was 0.9719. This means that the MLR model explains, or captures, 97 percent of the variation in the response variable. This is an extremely good fit; however, we need to do some multicollinearity checking. If there are some highly correlated predictors, this could be the reason that our R-squared value is so high.

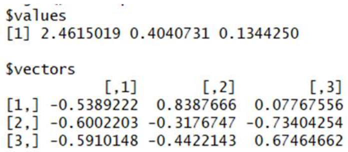
To start checking for multicollinearity between predictors, I unit length scaled my predictors so that each predictor vector had a mean of 0 and a variance of 1. Then, I looked at the correlation matrix (see *Figure 2*).



```
                mean_rating mean_share viewers_premier
mean_rating       1.0000000  0.6808953       0.6411832
mean_share        0.6808953  1.0000000       0.8633853
viewers_premier   0.6411832  0.8633853       1.0000000
```

Figure 2. Correlation Matrix for predictors in the MLR, generated in R.

The correlation between premier viewers and the mean share is fairly high, but not close enough to 1 to raise a lot of concern. Looking at the eigen decomposition of this unit length scaled correlation matrix can further help us identify if there is any multicollinearity between predictors (*Figure 3*).

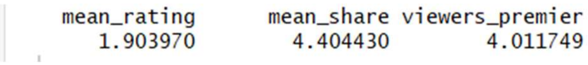


```
$values
[1] 2.4615019 0.4040731 0.1344250

$vectors
           [,1]       [,2]        [,3]
[1,] -0.5389222  0.8387666  0.07767556
[2,] -0.6002203 -0.3176747 -0.73404254
[3,] -0.5910148 -0.4422143  0.67464662
```

Figure 3. Eigen decomposition of the matrix in *Figure 2*, generated in R.

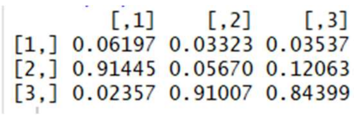
The lowest eigenvalue is the eigenvalue corresponding to premier viewers. In the premier viewers eigenvector, we can see that eigenvector 3, components 2 and 3, are highly correlated (meaning premier viewers and mean share are highly correlated) because they are relatively big and add up to a number that is close to 0. The max eigenvalue divided by the min eigenvalue, or the condition number, equals 2.461/0.134 or 18. Any condition number over 100 surely indicates multicollinearity. 18 is relatively high but still not high enough to conclude that multicollinearity exists between predictors premier viewers and mean share. We can look at the combined effect of the predictors dependence on one another by looking at the Variance Inflation Factors (VIF). The individual VIFs are a measure of how much the variances of each predictor are inflated by every other predictor being in the MLR model (see *Figure 4*).



```
 mean_rating      mean_share viewers_premier
    1.903970        4.404430        4.011749
```

Figure 4. The VIFs of each predictor in the MLR model, generated in R.

The VIFs are nothing but the diagonal elements of the inverse of the matrix in *Figure 2.* Since that matrix is unit length scaled, we can say that if any singular VIF is greater than 10 (some constant), then we have a significant VIF which indicates multicollinearity. All the VIFs in *Figure 4* are less than 10, so there is no indication of multicollinearity here. Lastly, we will look at the variance proportion matrix (see *Figure 5*).



```
        [,1]    [,2]    [,3]
[1,] 0.06197 0.03323 0.03537
[2,] 0.91445 0.05670 0.12063
[3,] 0.02357 0.91007 0.84399
```

Figure 5. Variance proportion matrix, generated in R.

This matrix explains the proportion of variance in a beta coefficient that is contributed by a predictor being in the model. For example, the value at row 3 and column 2 of the matrix in *Figure 5* indicates that 91 percent of the variation in the coefficient on predictor 2 (mean share) is caused by predictor 3 (premier viewers) being in the MLR model. Any proportion greater than 0.5 is taken to be significant, so *Figure 5* shows that multicollinearity exists in the MLR model.

In order to deal with this issue of multicollinearity, we will first try to fit a Ridge Regression model and examine the ridge trace plot (*figure 6*).

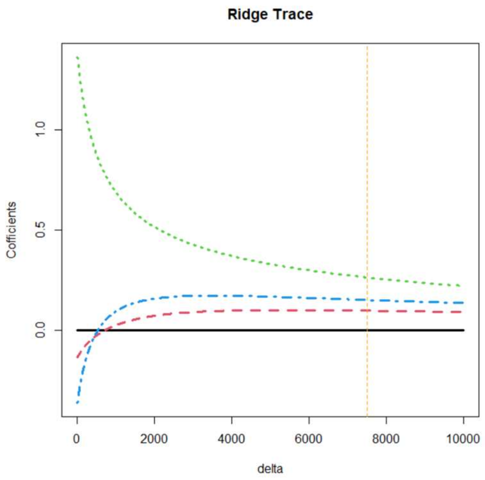


Figure 6. Ridge trace plot, generated in R

Figure 6 shows the Ridge Trace Plot for the full MLR model. Basically, we are adding delta values to the eigenvalues of each predictor to make the predictors more orthogonal with each other. As the predictors become more orthogonal, there is less variation in the coefficient estimates on each predictor. We can see that the coefficients are shrinking to more of a stable value as delta increases in the plot. The delta values are scaled up by 100 in the graph, so the vertical line is at delta = 75. As delta increases, the bias of coefficients on predictors increases and the R squared value for the Ridge Regression model decreases. Since delta is so large in this case, Ridge Regression is not a viable alternative to eliminate multicollinearity in the MLR model.

Instead of using Ridge Regression, I tried to fit a Principal Component Regression (PCR) model, which makes the predictors orthogonal with one another. I used the first two principal components which accounted for 95 percent of the variation in the response. The normal probability plot of the residuals is very heavy tailed, and the residual versus fitted values plot for the PCR model has some fanning in it. (see *Figure 7*).

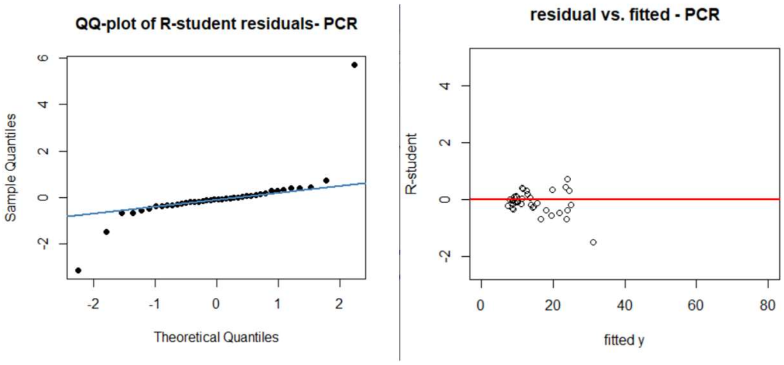
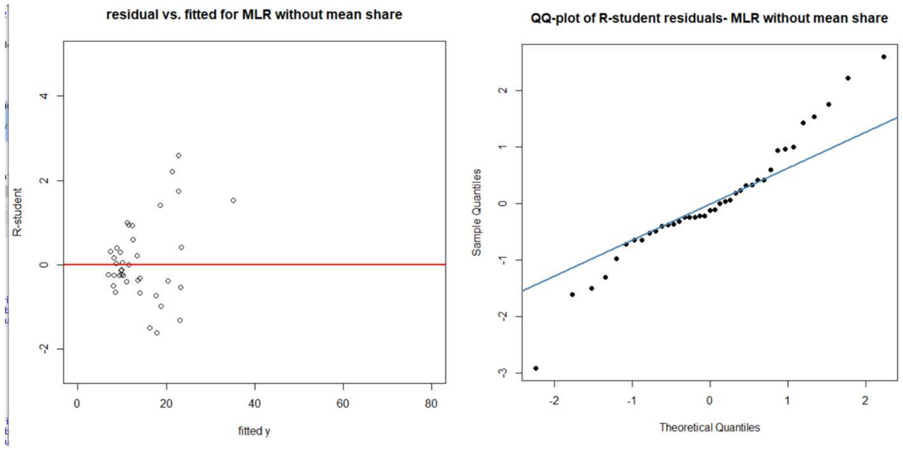


Figure 7. Residual Plots for the PCR model, generated in R

The PCR model violates some modeling assumptions from looking at *Figure 7*. If we use a different sample set of data, we could have a completely different model fit. The normal probability plot should have points that fall along the line, and the student residual vs fitted value plot should have random errors above and below the red lines. So, the best alternative to deal with the multicollinearity in the MLR model is to get rid of either predictor mean shares or premier views.

I selected the variable to get rid of that has a less significant impact on the MLR model, which is mean shares (even though it is a very marginal difference in significance between mean shares and viewers premier). Influential points analysis using this new MLR model without the predictor mean shares brings out an outlier we need to remove. The standardized residual for point 1 is over 5, which indicates that point 1 is an outlier in the new MLR model fit. Point one is also classified as influential because it has a cook’s distance of 0.3, so it is affecting the fitted values in the new MLR model the most. Further, DFFIT analysis confirms the cook’s distance result and shows that deleting point one will affect the fitted values by over 3 standard deviations, a huge amount. I deleted point one and examined the residual plots of the new MLR model (see *Figure 8*).



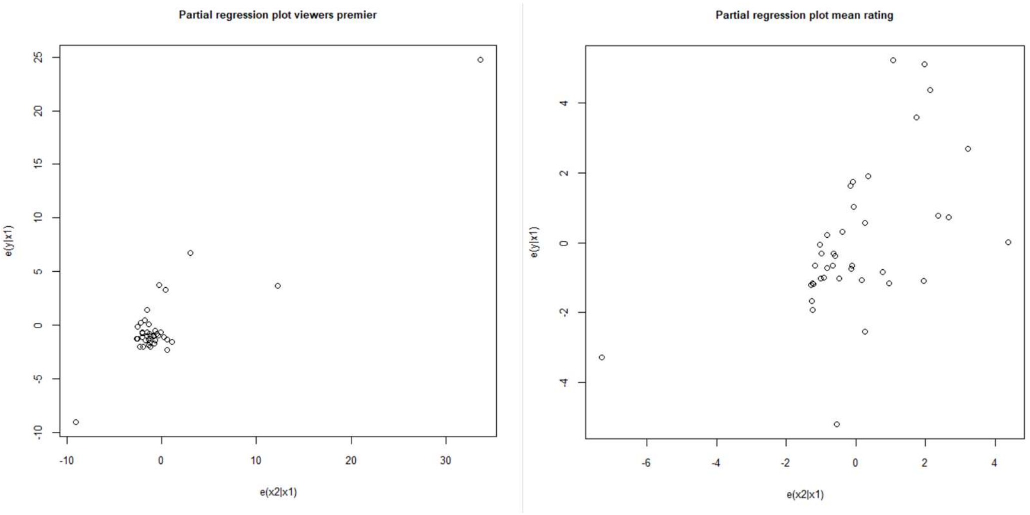
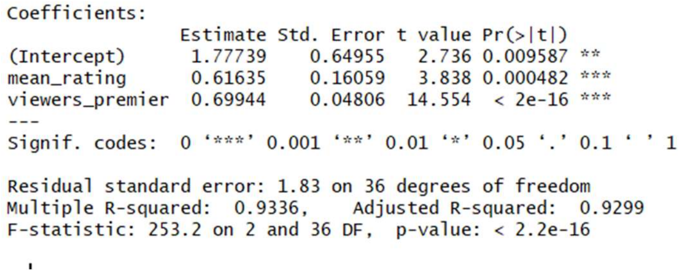


Figure 8. Residual plots of the new MLR model, generated with R.

Each partial regression plot in the bottom row of charts in *Figure 8* shows a slight linear trend, which is what we are looking for. The normal probability plot is slightly light tailed, but more normal than heavy tailed. The errors in the fitted values versus r student residuals are random and show no pattern. Overall, this new MLR model reduced the negative effects of multicollinearity and verifies all of the regression assumptions after we have deleted a negative influential point in the data. Here is the summary of our final MLR model for “Survivor” viewership (see *Figure 9).*


```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.77739    0.64955   2.736 0.009587 **
mean_rating      0.61635    0.16059   3.838 0.000482 ***
viewers_premier  0.69944    0.04806  14.554  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.83 on 36 degrees of freedom
Multiple R-squared:  0.9336,    Adjusted R-squared:  0.9299
F-statistic: 253.2 on 2 and 36 DF,  p-value: < 2.2e-16
```

Figure 9. Summary of the completed MLR model, generated with R.

The R squared value has decreased from 0.97 to 0.93, but our lack of describing the variation of the response is made up for by reducing the multicollinearity that existed when all variables were part of the MLR model. The predictors mean rating and premier viewers are significant when included in model. The final model equation is: finale viewers = 1.7739 + 0.6135*rating + 0.69944 * premier viewers. When either predictor is held constant and the other is increased by 1 unit, the number of finale viewers will increase by around 600 thousand people.

Conclusion

At the start of my project, I wondered if the rating method used to rate the episodes of “Survivor” could be trusted (*does a good rating reflect the number of finale viewers being higher*?). I also wondered if the number of season premier viewers was significantly tied to the number of season finale viewers. The MLR model I created in the analysis section gave me an answer to both of these questions. Since the beta coefficient on the rating predictor in the MLR model equation was significant, I can conclude that the rating is significantly tied to the number of finale viewers. Because the coefficient is also positive, a higher rating would reflect the number of finale viewers being higher; therefore, the rating system of the “Survivor” show is a viable system. Regarding second inquiry, the beta coefficient on the premier viewer predictor in the MLR model equation was significant, meaning that the number of season premier viewers was in fact significantly tied to the number of season finale viewers. The coefficient value was

0.6994, which means that for every 1 million additional premier viewers, there will be 699,400 more season finale viewers. Basically, "Survivor" retains 69 percent viewers each of the seasons from start (premier) to finish (finale). I was able to improve the model quite a bit by eliminating a predictor and removing some multicollinearity, but some still may exist because the R squared value is still quite high. The model coefficients are subject to a lot of change when new data is recorded if there is still multicollinearity; however, I have made a sufficient model that answers my questions for data recorded on "Survivor" seasons 1-40.

Sources

Data retrieval and idea

Rfordatascience. (n.d.). *Tidytuesday/readme.md at master · rfordatascience/tidytuesday*. GitHub. Retrieved December 15, 2021, from https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-06-01/readme.md#viewerscsv

Data set author

Oehm, D. (2021, May 1). *Survivor: Data from the TV series in R - daniel oehm: Gradient descending*. Daniel Oehm | Gradient Descending. Retrieved December 16, 2021, from http://gradientdescending.com/survivor-data-from-the-tv-series-in-r/

Code

All code and code ideas come from lecture slide created by the Regression Analysis Course professor Dr. Cheng Han Yu