Does Persistence Pay Off on Twitch?

An Analysis of Twitch Streaming Data

by Shawn Kim

Table of Contents

Introduction

Gaming has taken off in the last 15 years, not just as a huge entertainment industry for those who play, but for those who watch as well. This has all been possible because of the advances in technology that have taken gaming from simple pastimes on the weekends to immersive and streamlined experiences online. Consequently, streaming content as a serious, full-time occupation has become viable and even profitable for many streamers.

More and more people are now aware of the amount of money that can come from growing a successful stream on a streaming platform such as Twitch. Of course, the idea of producing content online for money is not new as everyone is familiar with online celebrities on famous platforms such as YouTube and Instagram. However, unlike these traditional platforms that emphasize “static” content, Twitch has set itself apart by providing a platform supporting real-time streaming of content where content creators and fans gain the ability to interact with each other in real-time. The sense of community and personal involvement that this creates for both viewers and streamers alike is the main appeal of live streaming over other forms of content.

Although streaming can be a lucrative career, not everyone has what it takes or the luck to reach the top. Many successful streamers have streamed for years with no recognition at the start, some have built their fame in a relatively short amount of time, and the vast majority of others have or will continue to stream with little to no success in sight. With the context in mind, this project attempts to answer the question of whether persistence pays off on Twitch by analyzing the streaming data for the top 1000 streamers on Twitch. Specifically, the total number of hours every Twitch channel has been on will be compared against other metrics such as average viewership per stream, the numbers of followers, and etc. Thus, the total hours streamed

should have a positive correlation with the various stream metrics if persistence does indeed pay off.

Methods

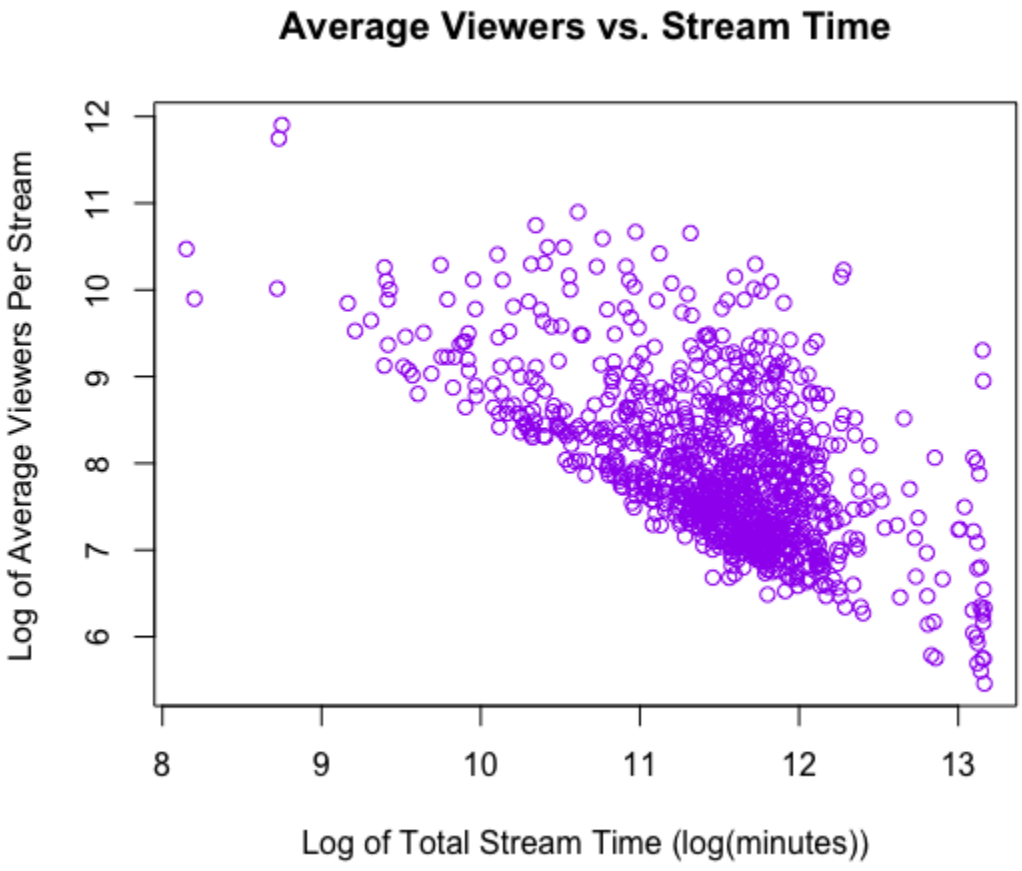
The Twitch stream data for the top 1000 streamers came from the online website Kaggle which provides free and comprehensive datasets for a diverse set of topics. As mentioned in the introduction, the explanatory variable we are using will be the total stream time for each channel. You can think of a Twitch channel like a YouTube channel with the distinction that a YouTube channel has videos while a Twitch channel can only have a stream that is on or off. The total stream time was chosen as the explanatory variable since we want to find out whether persistence plays a role in streaming success.

The response variables are then average viewers per stream (meaning average viewers for the duration of the time a stream is on), peak viewers per stream, watch time per stream (watch time meaning the cumulative sum of the watch time of every individual), followers (analogous to Subscribe button on YouTube), and followers gained (since the time that a Twitch channel was first created). These response variables are chosen to quantify the success of Twitch channels but certainly are not all the available Twitch metrics that can paint the full picture. I picked these variables because they can easily be understood at face value without getting into Twitch specifics.

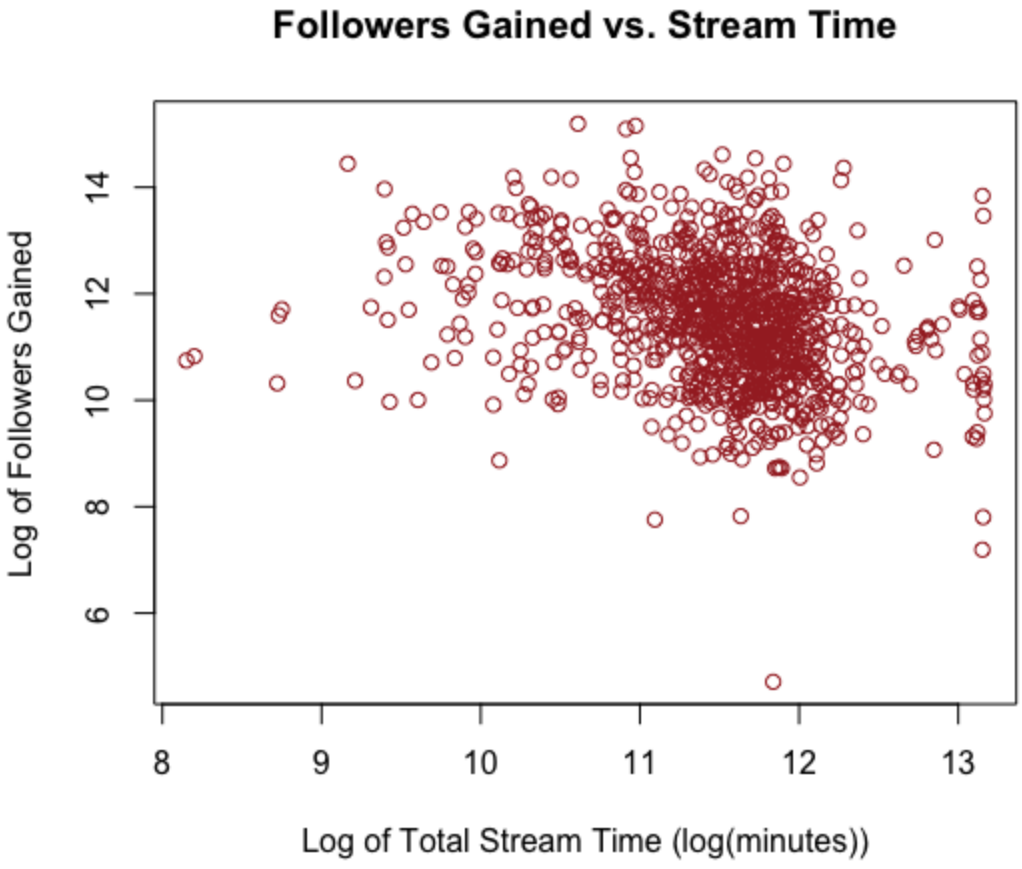
A shortcoming of the dataset used for this project is that it is only for the top 1000 streamers on the Twitch platform which under normal circumstances would be extremely biased towards those that already found success when considering whether persistence pays off. However, it is important to note that viewership traffic on Twitch is extremely disproportionate

towards the top end and that the top 1000 already includes a very large portion of streams that only average a couple hundred or less viewers per stream in contrast to the top 50 streams that will average in the tens of thousands to hundreds of thousands. To consider the other thousands of registered Twitch channels that get 0 views average per stream and have minimal overall stream time anyways would skew all the data to irrelevancy. Rather, looking at the top 1000 channels keeps the data relevant to streams that you might actually find on the Twitch platform and strikes a balance between the very top end and a good portion of streamers that get modest or acceptable viewership traffic.

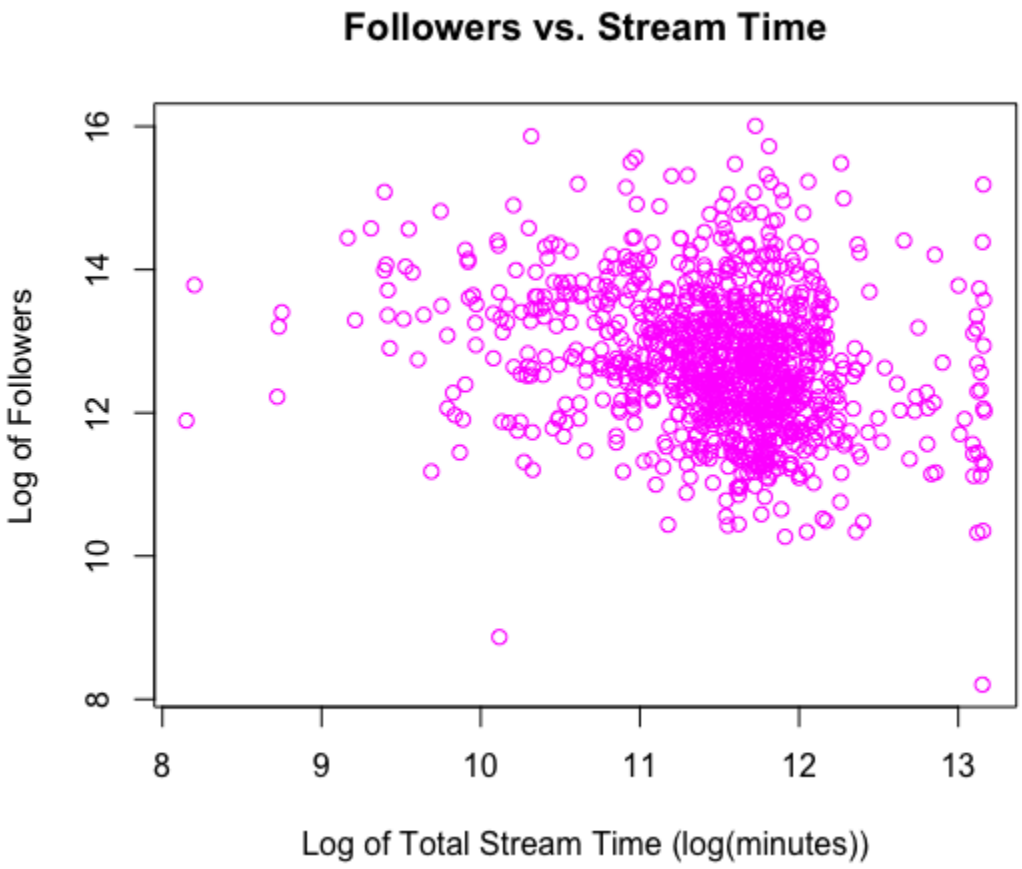
Results



This scatter plot shows the log of the average viewers per stream versus log of the total stream time. We can see that there is a negative correlation between the log of the explanatory and response variables in this case. Therefore, increasing log of total stream time generally tends to decrease log of average viewership per stream. This suggests perhaps that stream success is more about the quality of the content rather than quantity of it. There are outliers at both ends of the spectrum. For the outliers that have a very high log of average viewership per stream compared to log of total stream time, these are representing company owned Twitch channels that only go live for their special events and tournaments which will attract a lot of viewers. On the other end where there is a very high log of total stream time compared to log of viewership average, these can be explained by channels that like to rerun pre-recorded 24/7 which will not garner a lot of attention.

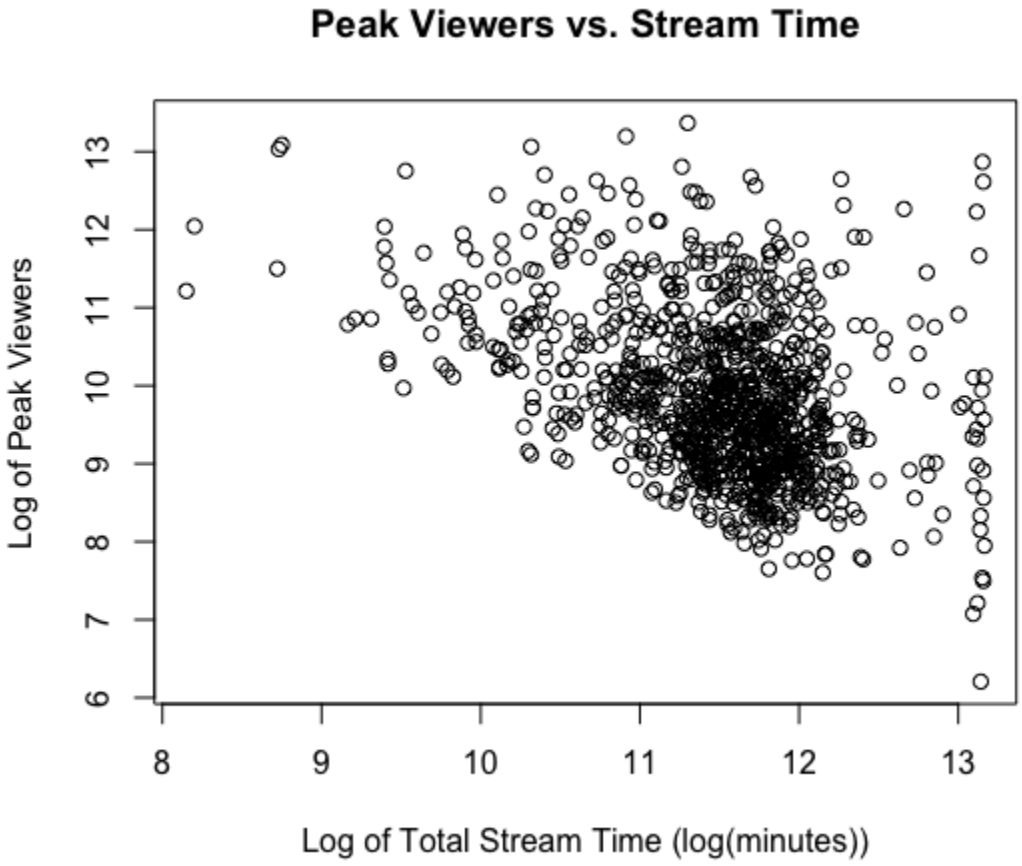


This scatter plot shows log of followers gained versus log of total stream time. Here, there either seems to be no correlation or a very slight negative correlation between the two variables. The followers gained over the lifetime of a Twitch channel will depend on a lot of other things than the total time the stream has been on. This again suggests that persistence is not the right key for streaming success.

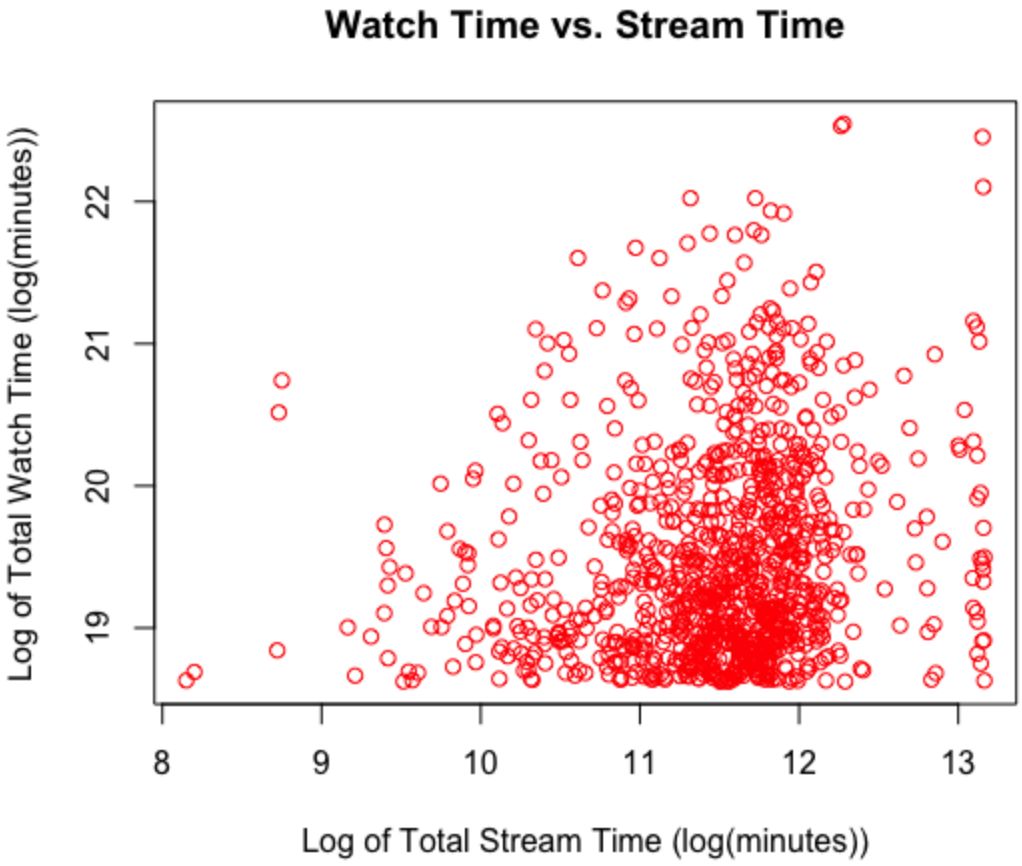


This scatter plot shows the log of followers versus log of total stream time. Again, the current amount of followers that a stream has does not necessarily have anything to do with the amount of time the stream has been on. The scatter plot is also very similar to the previous plot

of followers gained, perhaps showing that people do not tend to unfollow streams that they have already followed.



This scatter plot is showing the log of peak viewers versus the log of total stream time. Similar to the first plot of the average viewers, there appears to be a negative correlation between these two variables. It makes sense that the plot of peak viewers is similar to the plot of average viewers per stream since average and peak are generally closely related to each other.



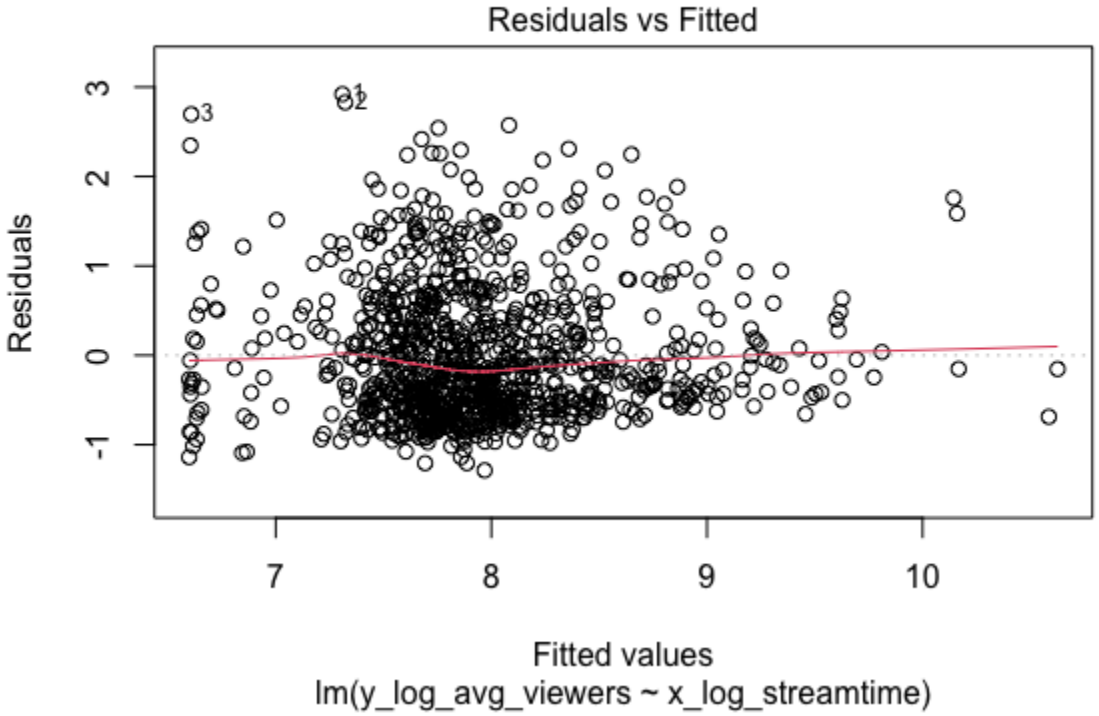
This scatter plot shows the log of the watch time versus the log of the total stream time. There may be a positive correlation between total watch time and total stream time since having the stream on more gives more opportunity for watch time. However, the plot is still very scattered overall since the number of people you have watching a stream to accumulate total watch time is not tied to how long you have your stream running all day. Therefore, this does not support persistence as an indicator of stream success.

Note: Please refer to the appendix for all R-codes related to plots. The csv file for the twitch data can be downloaded from Kaggle cited in the references.

Analysis

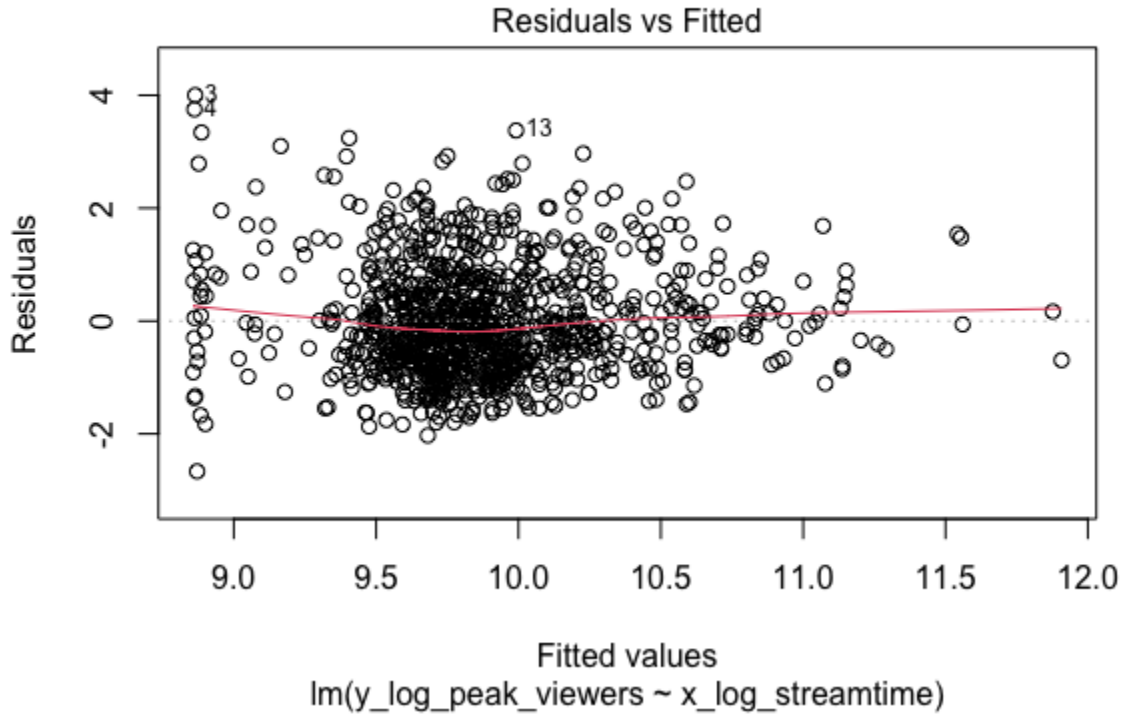
Linear regression models were constructed using R in order to visually indicate the degree of normality and/or linearity of the log-log graphs. The residuals versus fitted plots were used to test for the linearity and the normal quantile-quantile plots were used to test for normality. The residuals versus fitted will be discussed first.

Residual vs. Fitted Plots



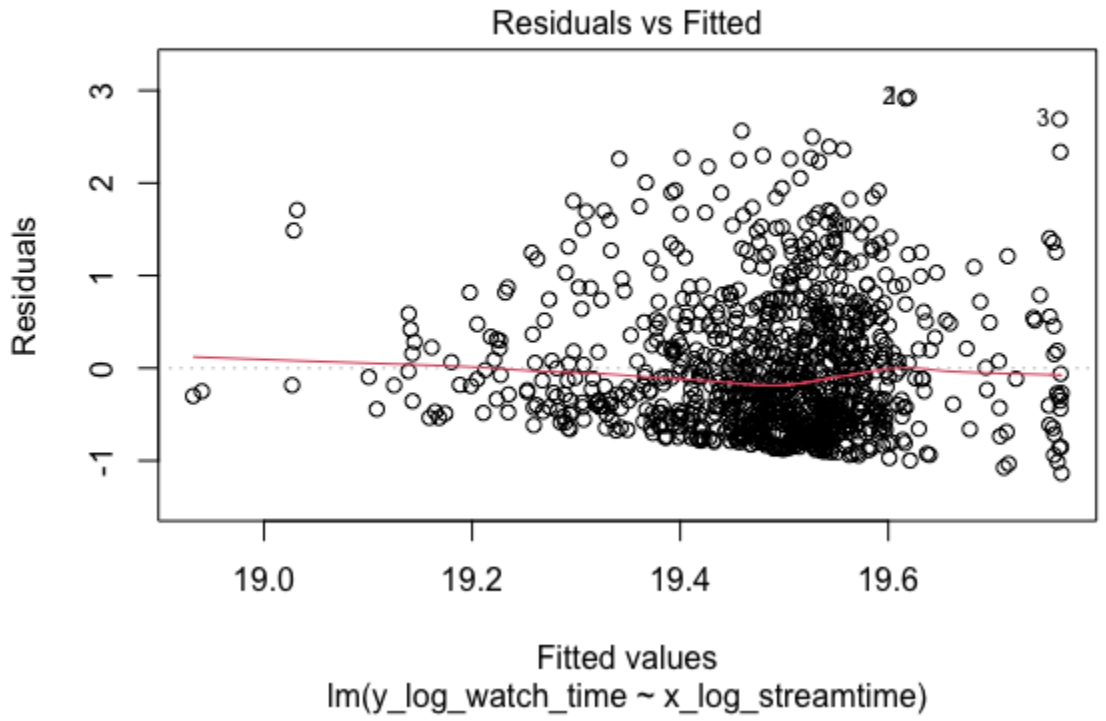
For the first plot of residuals versus fitted for the log-log graph of average viewers versus stream time, we can see that most of the data points are clustered towards one spot. Linearity is indicated by an even spread distribution of data points along the line so what we have visually indicates that there is not a strong case for linearity in the relationship between log of average versus log of stream time.

This plot also indicates the non-constant variance in the data as shown by the vertical spread of points along the dotted gray line (the horizontal line) that gets smaller in range when going left and right from the main cluster of points. Additionally, we see that some residuals clearly stand out from the others, indicating that there are some outliers in the data set.

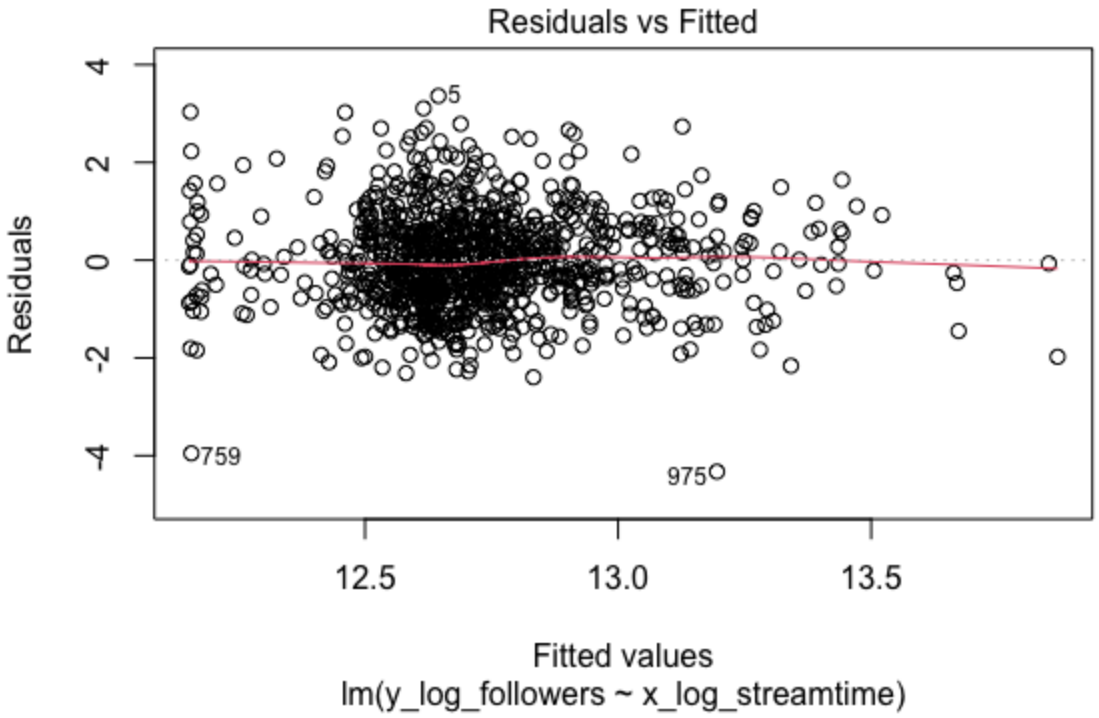


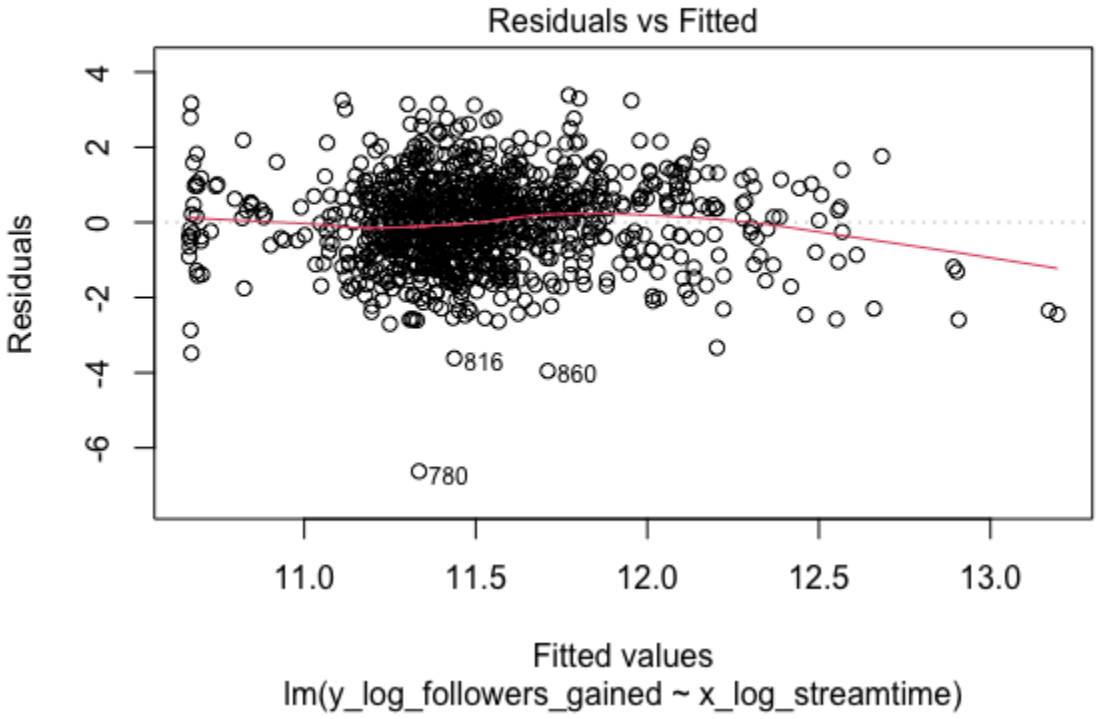
Next, we have our residuals versus fitted values for our log of peak viewers versus log of stream time. Much like our previous residual plot of average viewers, we see that the data points

are clustered together and not evenly spread out along the horizontal line which indicates non-linearity. The uneven distribution of data points along the vertical access also indicates non-constant variance in the data, and again we see that there are outliers in the data set.



For our residuals versus fitted plot of log of watch time, we again see the case as with the previous two graphs. The spread of points along the horizontal line indicates non-linearity, non-constant variance, and some extreme outliers.

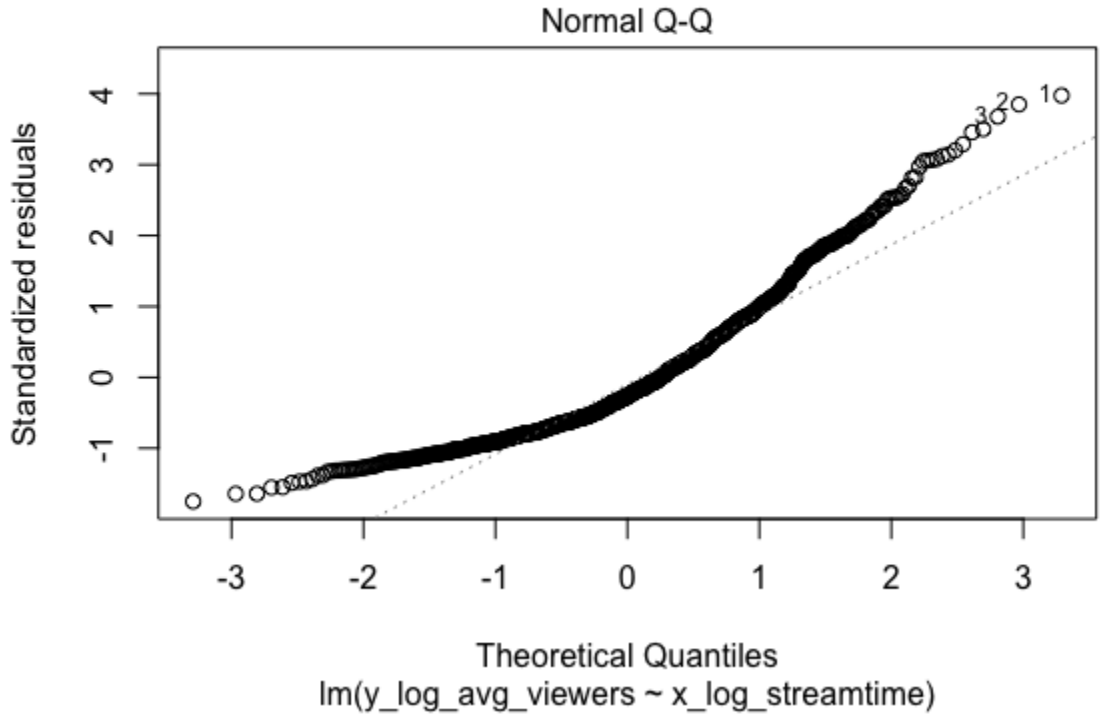

Residuals vs Fitted
Residuals
Fitted values
lm(y_log_followers ~ x_log_streamtime)
5
759
975
12.5
13.0
13.5
4
2
0
-2
-4

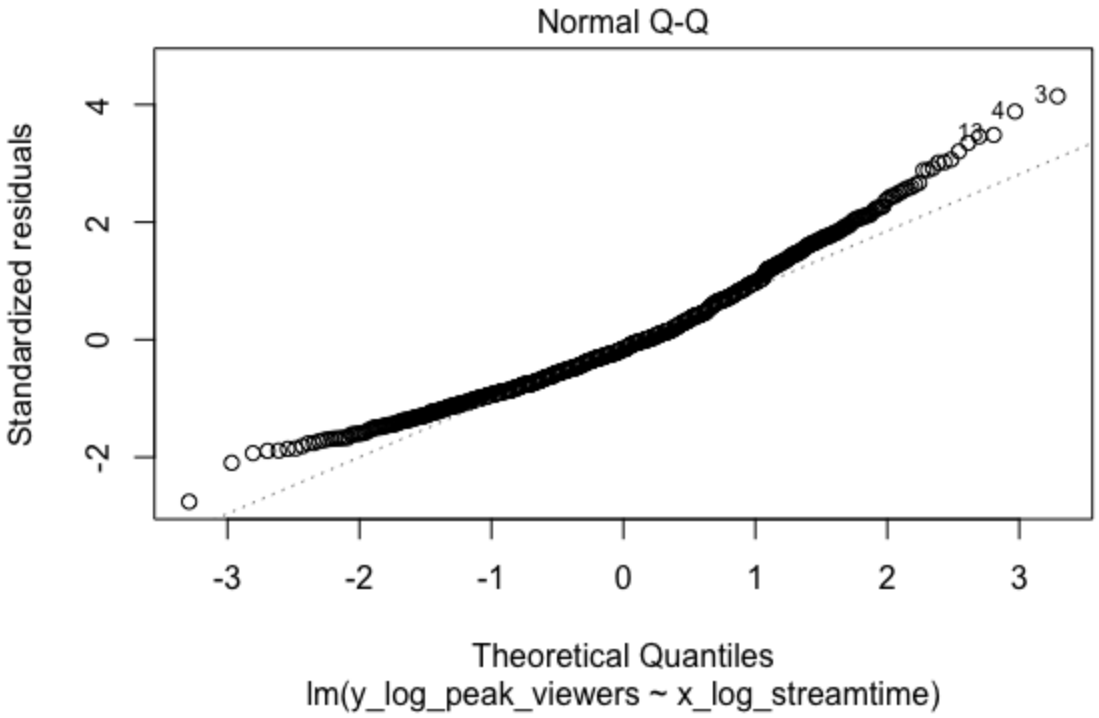

Residuals vs Fitted
Residuals
Fitted values
lm(y_log_followers_gained ~ x_log_streamtime)
816
860
780
11.0
11.5
12.0
12.5
13.0
4
2
0
-2
-4
-6

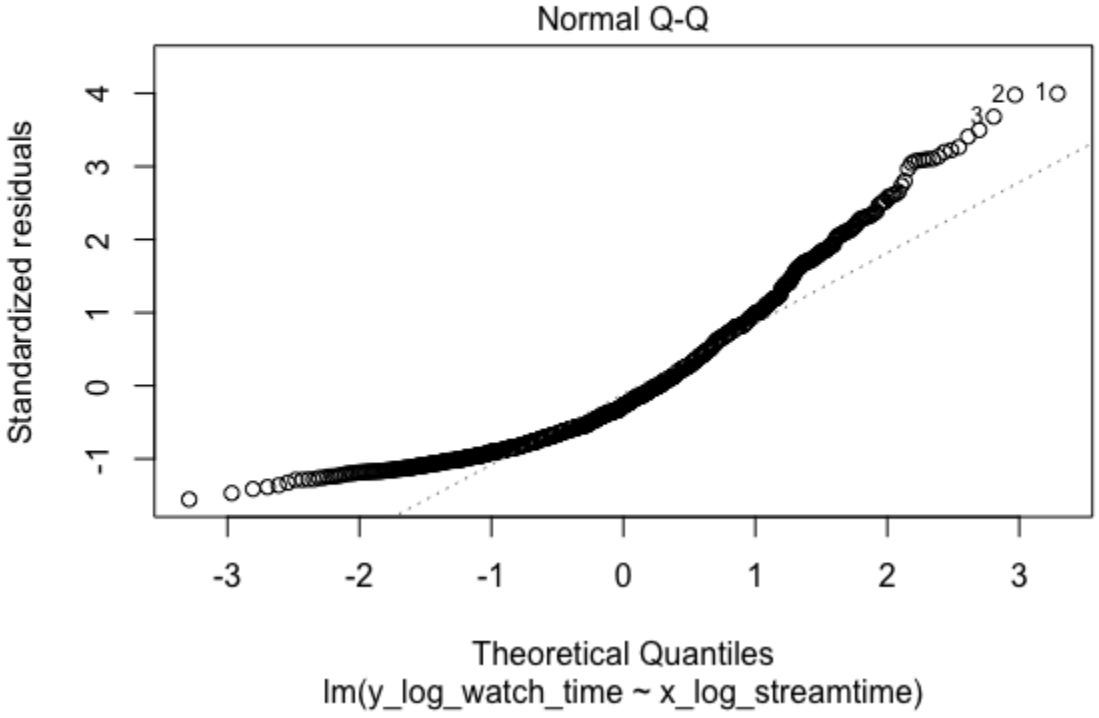
For the last two residuals versus fitted of the log of followers and the log of followers gained, we see that they are again a similar case to the first three plots in terms of the clustering of residuals and the presence of clear outliers in the data set.

However, these two plots are different in that the data points are much more evenly distributed on both sides of the horizontal line, indicating that the log-log plots of these data are much more linear than the first three that were introduced. Additionally, we can see that the spread across the line is not only more even than the first three graphs but also smaller in range, indicating less overall variance in the data set although the variance itself is still non-constant. Next, the normal quantile-quantile plots will be examined.

Normal Q-Q Plots

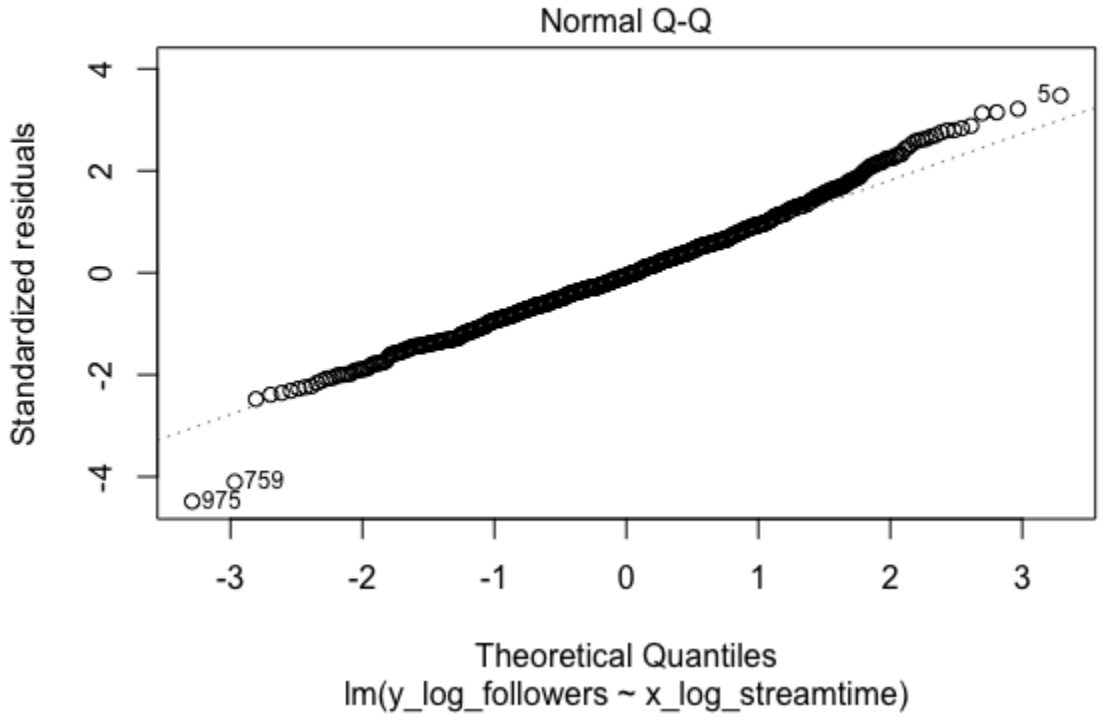


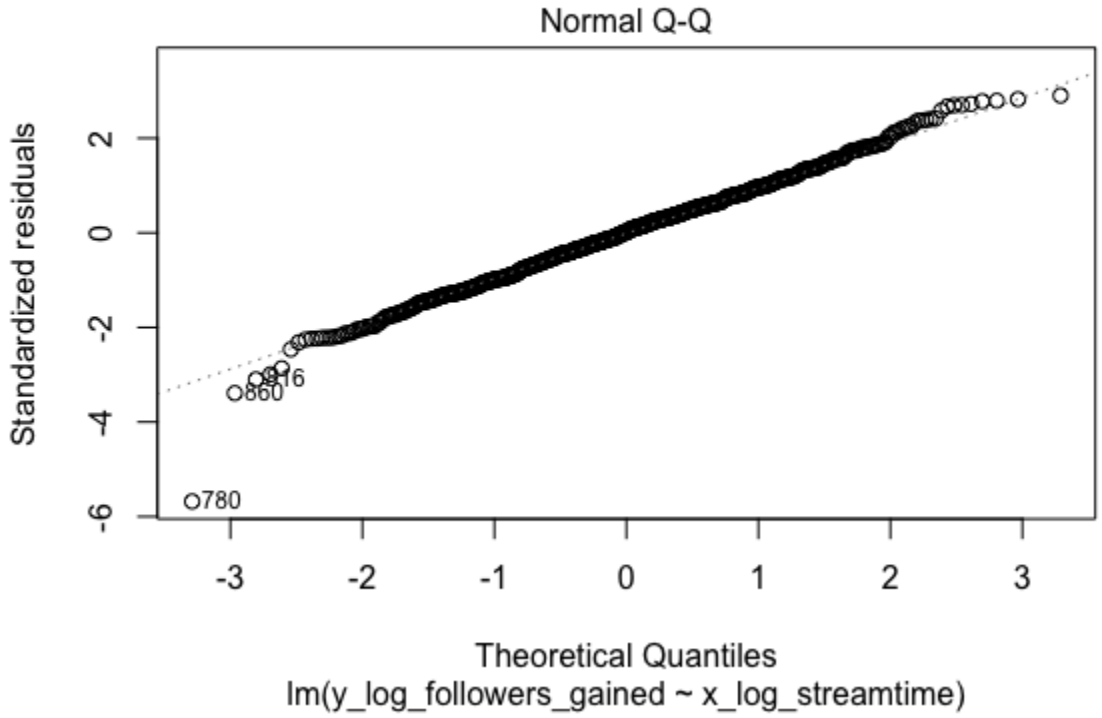

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y_log_peak_viewers ~ x_log_streamtime)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y_log_watch_time ~ x_log_streamtime)

The first normal Q-Q plots shown are for the log of the average viewers, log of the peak viewers, and log of the watch time respectively (we are going in the order they were shown for the residual plots). Normality is indicated by how well the data points fall along a straight line (in this case, how long they fall along the dotted gray line). For these first three plots, we can see that they do not fall well along the straight line and opt for more of a curve instead, telling us that the log-log data sets are most likely not normally distributed.

The curvature in our Q-Q plots also tell us about the kind of skewness our data set has. A curvature upwards indicates right skewness and the degree of curvature upwards also reflects the degree of right skewness. We can see that the log of peak viewers versus the log of stream time has the most normally distributed data out of the three. Next, we'll take a look at the Q-Q plots for the log of followers and log of followers gained respectively.





We can see that these last two plots fall very closely along the straight line, indicating that the data sets for log of followers and log of followers gained versus log of stream time come from a normally distributed population. Some other things to note are that both graphs have some notable extreme outliers and that the Q-Q plot of log of followers seems to have a slight right skew while the Q-Q plot of log of followers gained seems to be curving slightly downwards at the end which indicates some degree of left skew due to the extreme outliers.

From examining the Q-Q plots alongside the log-log plots of the data, I tried to make sense of why the first three Q-Q plots indicated a right skewed distribution for their log-log data plots as opposed to the indication of non-skewness of the last two plots. My best guess and

intuitive understanding is that if I took random samples of data points from any of the log-log plots and plotted the response variable (such as average viewers) as the x-axis, and the rate of occurrence of the response variable as the y-axis, I would have ended up with distribution curves that matched the skewness indication from the normal Q-Q plots. I think this because I observe that the spread of data points along any vertical line for the log-log plots is clustered at a specific range and becomes more spread out as we go away from this range in both directions. The clustering of these data points for each log-log plot is consistent with what I would expect the distribution curve to look like if I had plotted it the way that I described for each log-log plot.

Conclusions

Based on the evidence so far, it appears that persistence does not have any apparent relationship to streaming success, at least if we are measuring persistence in terms of the hours someone has streamed. In hindsight, something as one dimensional as stream time can not be equated to a broad word such as “persistence”, and that may also be the pitfall of this investigation since being one dimensional is the point of an explanatory or response variable in statistics. I do still believe that persistence does play a key role in streaming success, but now have the understanding that it may be more about what is happening off stream rather than on stream.

But on the topic of “on stream”, I have an explanation of why the data did not turn out to be very insightful in this case. First, it is important to note that just like with Youtube, Twitch channels do not have to be owned by a single person and can be owned by a group of people, an organization, or a company. Companies have a lot more publicity than individual people and tend to go live only to host major events that will attract the attention of many viewers. These

company channels account for the data points that represent very high average/peak viewers and number of followers with relatively low overall streaming time. The dense cluster of points found in every plot represents where most of the top 1000 streams are in terms of viewership and followers compared to stream time and most of these channels will be solo content creators, much like the majority of Youtube channels. Finally on the other end are channels that have a very high relative total stream and low viewership and followers. These channels are most likely ones that rerun pre-recorded content 24/7, thus defeating the purpose of a live stream and consequently not attracting much attention in the long run. In summary, there were a lot of factors and qualitative details with the Twitch channels that were not accounted for by this experiment. Not accounting for these hidden variables is the reason why the results were contrary to the expectation of a positive correlation between the explanatory variable of total stream time and the response variables of the other Twitch stream metrics used to indicate success.

In terms of the usability of this study, it is not very valuable as it is now since the log-log data did not show any good indications of linearity in the end and the suggested normality of the log-log data could not be clearly seen from the plots either. However, I believe that there is good room for improvement if I were to do this again or if someone had a similar idea to this. Instead of looking for a linear correlation between two variables (as we have already discovered that there are too many hidden variables for this), we should sample the occurrences of the response variables from the data set. An example could be to sample channel viewership from the data set to plot as a distribution and see if the sample mean of this distribution changes from year to year.

References

Mishra, A. (2020, August 24). Top streamers on twitch. Kaggle. Retrieved November 18, 2022, from https://www.kaggle.com/datasets/aayushmishra1512/twitchdata

Appendix

```
twitch <- read.csv(file=/path/to/directory/twitchdata-update.csv)

#log average viewers vs log stream time
plot(x=log(twitch$Stream.time.minutes.),
y=log(twitch$Average.viewers), xlab="Log of Total Stream Time
(log(minutes))", ylab="Log of Average Viewers Per Stream",
col="purple", main="Average Viewers vs. Stream Time")

#log peak viewers vs log stream time
plot(x=log(twitch$Stream.time.minutes.), y=log(twitch$Peak.viewers),
xlab="Log of Total Stream Time (log(minutes))", ylab="Log of Peak
Viewers", col="black", main="Peak Viewers vs. Stream Time")

#log watch time vs log stream time
plot(x=log(twitch$Stream.time.minutes.),
y=log(twitch$Watch.time.Minutes.), xlab="Log of Total Stream Time
(log(minutes))", ylab="Log of Total Watch Time (log(minutes))",
col="red", main="Watch Time vs. Stream Time")

#log followers vs log stream time
plot(x=log(twitch$Stream.time.minutes.), y=log(twitch$Followers),
xlab="Log of Total Stream Time (log(minutes))", ylab="Log of
Followers", col="magenta", main="Followers vs. Stream Time")

#log followers gained vs log stream time
plot(x=log(twitch$Stream.time.minutes.),
y=log(twitch$Followers.gained), xlab="Log of Total Stream Time
```

```
(log(minutes))", ylab="Log of Followers Gained", col="brown",
main="Followers Gained vs. Stream Time")

#storing x and y of log of data
x_log_streamtime = log(twitch$Stream.time.minutes.)

y_log_avg_viewers = log(twitch$Average.viewers)

y_log_peak_viewers = log(twitch$Peak.viewers)

y_log_watch_time = log(twitch$Watch.time.Minutes.)

y_log_followers = log(twitch$Followers)

y_log_followers_gained = log(twitch$Followers.gained)

#linear modeling
avg_viewers_lm = lm(y_log_avg_viewers~x_log_streamtime,data=twitch)

avg_peak_viewers_lm=lm(y_log_peak_viewers~x_log_streamtime,data=twitc
h)

avg_watch_time_lm=lm(y_log_watch_time~x_log_streamtime,data=twitch)

avg_followers_lm=lm(y_log_followers~x_log_streamtime,data=twitch)

avg_followers_gained_lm=lm(y_log_followers_gained~x_log_streamtime,da
ta=twitch)

#plotting linear model graphs (hit enter for all prompts)
plot(avg_viewers_lm)

plot(avg_peak_viewers_lm)

plot(avg_watch_time_lm)

plot(avg_followers_gained_lm)

plot(avg_followers_lm)
```