Social Media and Art Patronage

STATS 767 Project

Saurabh Gupta (ID: 567 20 2721)

23 October 2018

Contents

1 Introduction

1.1 Effect of Social Media on Art Patronage

Research Question: *Is the social media presence of creators (artists) related to their online patrons?*

Patreon.com is an online platform that allows creators to seek patrons (similar to paid subscribers) pledging to pay as little as $1 per month to more than $1000. The payments and the distribution technology is taken care of by Patreon in exchange for a 5% fee (and some charges). It hosts more than 122 thousand creators and 4 million pledges.

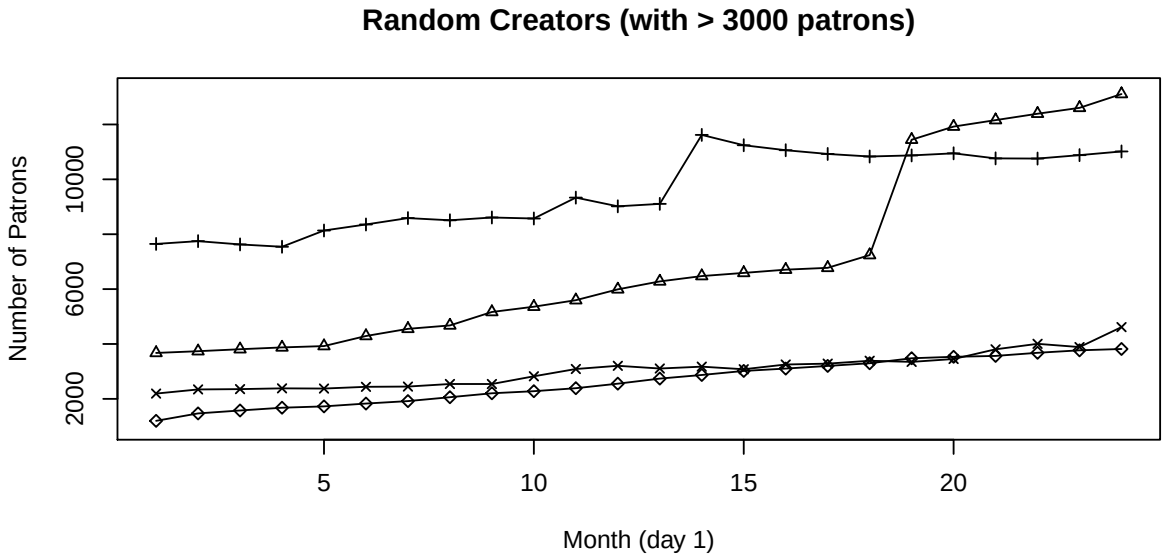
1.2 The dataset

Time series data on the number of `Patrons` and 5 social media statistics for top 103 creators in different categories. It has 103 rows wherein each creator is an observation. There are 138 variables (columns) that include 23 months' data for 6 variables. Social media metrics include Facebook Likes, Twitter Followers, YouTube Subscribers, YouTube Videos and YouTube Views. Data is measured on the 1st of each month and the time series window is from 1st July 2016 to 1st May 2018.

1.3 Plot of Number of Patrons

Figure 1 plots the monthly number of `Patrons` for 10 randomly chosen creators. It inidicates a trend in the data. Plots of other variables, not included here, also indicate a somewhat quadratic trend.

[0]Notes: Brand and creator names used in this presentation are copyrights of their respective owners. The data has been sourced from https://www.patreon.com/ and https://graphtreon.com/. It cannot not be published without their permission.



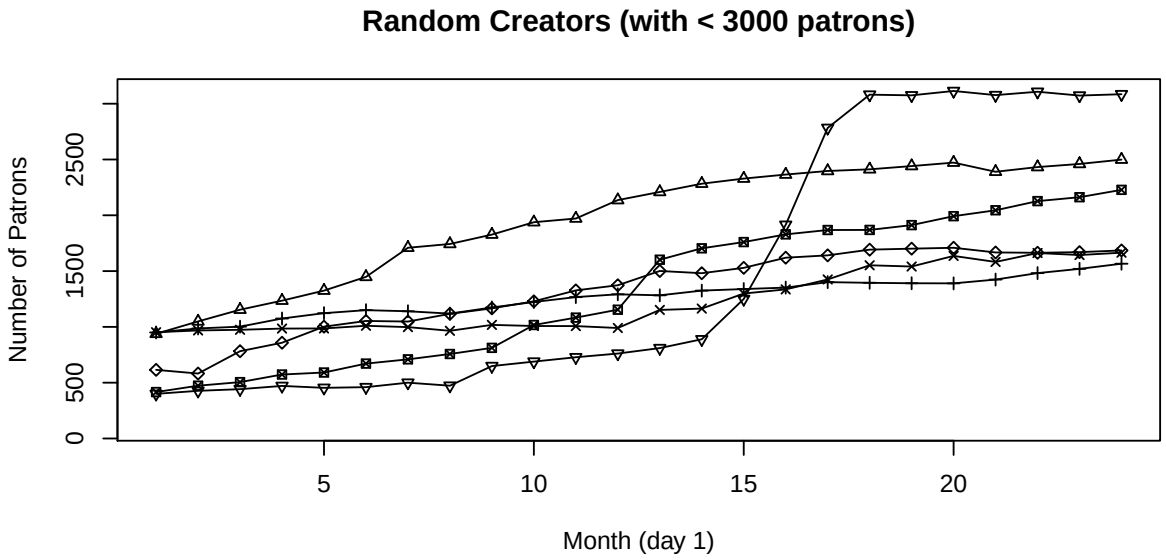


Figure 1: Number of Patrons for 10 random creators

2 Methodology and Assumptions

The trend in all variables can result in spurious correlations. To take a closer and unbiased look, we may detrend the data by regressing each of the variables on time period $t = (1, 2, \ldots, 23)$.

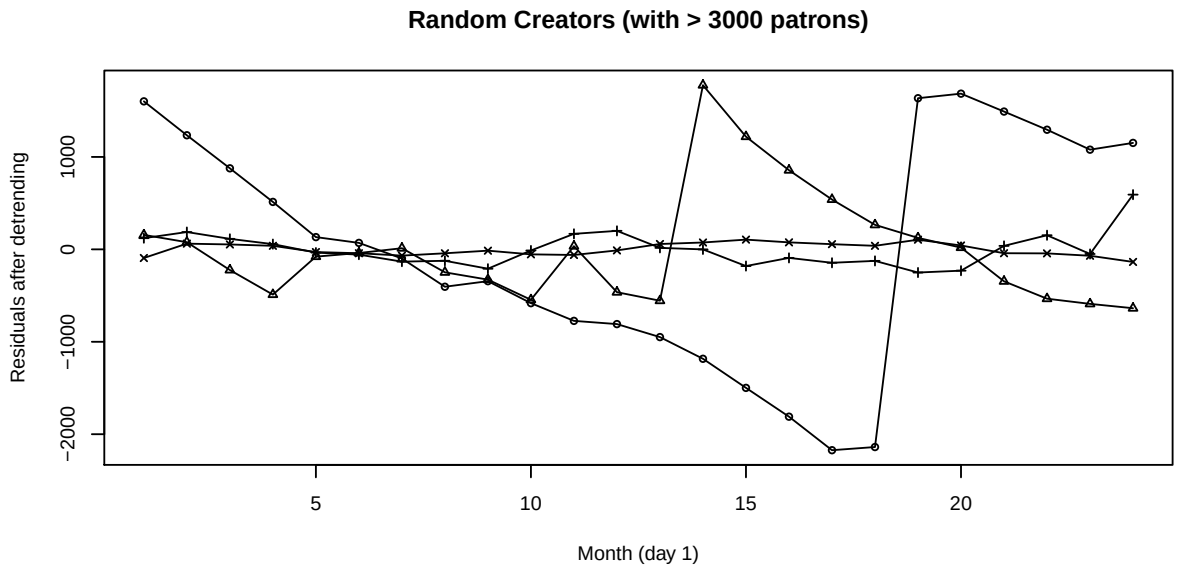
The number of `Patrons` aren't expected to increase indefinitely. Hence, I used quadratic function of time to detrend the data. After detrending, we are left with the residuals. They represent the variability in the data.

2.1 Plot of Patrons after quadratic detrending

Figure 2 plots the variable `Patrons` after their quadratic detrending. The data seems more mean stationary i.e. after some variability, it appears to be returning to its mean. Hence, we may naively assume that its mean doesn't depend on time t.

The number of variables, $p = 138$ is larger than the number of observations, $n = 103$. Hence, I will perform PLS Correlation to check the relationship between variability in social media metrics and the number of `Patrons`.

- Let, Y denote the variability in `Patrons` for each of the 23 months.
- X denotes the the variability in social media metrics for the same period.



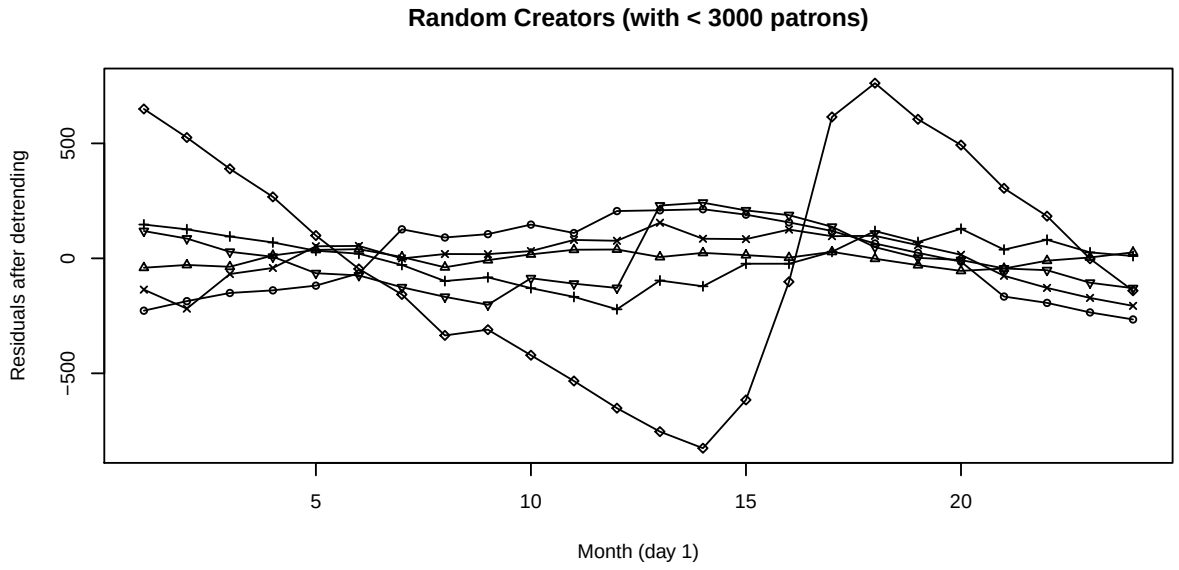


Figure 2: Patrons after quadratic detrending

2.2 Assumptions

First we may assume that measurement errors and biases are acceptable. We expect this sample of data to be adequately representative of the population.

Second, we assumed there was a trend in the data after looking at the plots. Hence, it was tested using a linear model with the time t and t^2 as regressors i.e. quadratic trend. Its results have been presented below as p-values of the coefficients acros 103 creators. To make the data comparable, quadratic detrending of X variables was done in a similar fashion before making the X matrix with the residuals.

Upon linearly regressing `Patrons` against the time variable, $t = 1, 2, \ldots, 24$ using **R** function `lm()`, the coefficient for linear term t was significant for 83 of 103 creators while the quadratic term t^2 was significant for only 60 of 103 creators, assuming a significance level of 0.05 for $Pr(> |t|)$. The number of `Patrons` or other social media variables aren't expected to increase indefinitely. Hence, I used quadratic function of time to detrend the data.

The p-values are summarised as quantiles in table 1.

Table 1: Significance, $Pr(> |t|)$, of coefficients for quadratic trend for 103 creators

Quantile	Intercept	Linear (t)	Quadratic (t^2)
0%	< 2.22e-16	< 2.22e-16	5.5310e-15
10%	< 2.22e-16	1.9575e-12	1.1369e-07
25%	< 2.22e-16	5.3524e-09	6.6676e-05
50%	3.849e-15	9.5293e-06	0.021071
75%	3.9475e-08	0.02072	0.340710
90%	0.073937	0.23427	0.710802
100%	0.837663	0.87047	0.995092

Third, the residuals after quadratic were used for PLS-Correlation. We may assume that the residuals are skewed Normal in their distribution. We can observe these in the Normal Quantile Quantile plots for some of the variables in figure 3 . PLS-Correlation is expected to be optimal with Normally distributed data. However it doesn't rely on parametric probability distributions. Hence, it is still appropriate for our analysis.

Finally, PLS correlation is based on linear combinations of the Y and X variables. Hence, if there were non-linear relationships in the data, they may not be observable in the results.

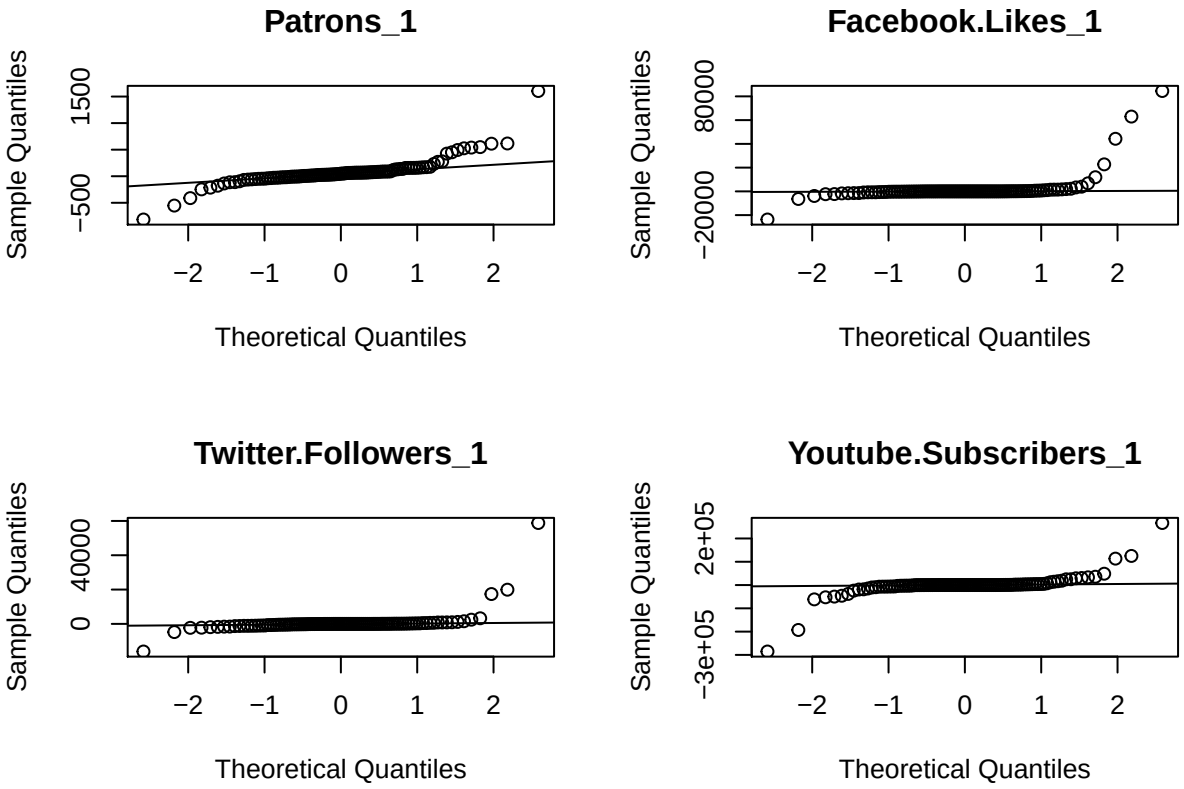


Figure 3: Normal QQ plots of Residuals after quadratic detrending of variables

2.3 Y and X variables for PLS correlation

The dimensions of **Y** variables is 103 rows and 23 columns. Rows are the creators while columns are the time series data of `Patrons`. The first 10 variable names indexed by month t has been provided below for reference.

```
## Dimensions of Y
```

```
## [1] 103  23
```

Names of first 10 Y Variables
Patrons_1
Patrons_2
Patrons_3
Patrons_4
Patrons_5
Patrons_6
Patrons_7
Patrons_8
Patrons_9
Patrons_10

The dimensions of **X** variables is 103 rows and 115 columns. Rows are the creators while columns are the 23 months' time series data of 5 social media metrics. The first 10 variable names indexed by month t has been provided below for reference.

```
## Dimensions of X
```

```
## [1] 103 115
```

Names of first 10 X variables
Facebook.Likes_1
Twitter.Followers_1
Youtube.Subscribers_1
Youtube.Videos_1
Youtube.Views_1
Facebook.Likes_2
Twitter.Followers_2
Youtube.Subscribers_2
Youtube.Videos_2
Youtube.Views_2

3 Partial Least Squares Correlation

3.1 Correlation between the canonical variates of Y and X

PLS Correlation gives the correlation between a linear combination of Y variables with that of X variables. It selects linear combinations that maximise the correlations within the variables. The first axis represents the the first pair of linear combinations that has the most correlation. The second axis is the another pair of linear combinations that has the second highest correlation.

The correlation results are then tested by finding correlations after randomly re-arranging rows of Y i.e. for different permutations of the data. If the original result is higher than a predefined percentage of permutations, we may say that the correlation is significant. For this study, we are assuming that if less than 5% permutations have a higher correlation, the result is significant. If more than 5% but less than 10% permutations have a higher correlation, then the results are weakly significant.

For this study, approximately 10-11% of 100 permutations have larger correlation than observed for the first axis and 11-12% for the second axis. This implies that there is weak evidence of a relationship between `Patrons` and social media metrics.

Note that scaling of variables is done by default in the `pls` function in **R**. It makes sense to scale because different metrics and different creators are at different scales.

Correlation between the first pair of canonical variates

```
## Observed Correlation: 0.63
```

```
## p-value of correlation based on 1000 permutations: 0.108
```

Correlation between the second pair

```
## Observed Correlation: 0.62
```

```
## p-value of correlation based on 1000 permutations: 0.114
```

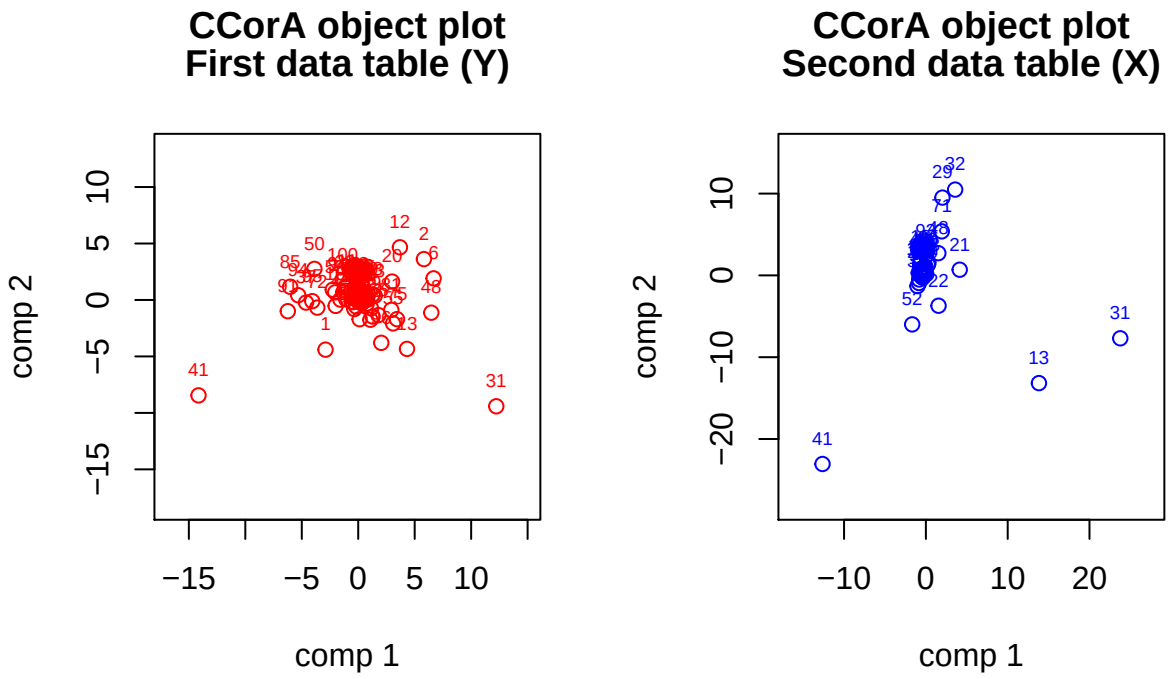


Figure 4: Biplot of Objects: plots the correlation of the observations (rows) with the first two PLS axes. On the left is the plot for PLS axes of Y while on the right is that for X.

3.2 Biplot

The biplots for objects and variables have been plot separately in figures 4 and 5, respectively, so that they are more visible. However, they appear on separate pages because of it.

3.3 Months well represented by PLS axes of Y

The biplot of variables for **Y** (left plot in figure 5) indicates:

- Months 5 to 8 have high positive correlation with axis 1. Going by the plots of residuals of `Patrons` in figure 2, these were the periods with apparently lower variability (not tested statistically).
- First 3 months and months 13 to 16 have high negative correlation with axis 1. Again, going by the plots, these were the periods with high variability.
- Axis 2 doesn't have high correlation with most months (except 14, 21).

This implies that PLS axis 1 indicates low or high variability in `Patrons`. However, PLS axis 2 isn't strongly correlated with most months, except month 21. There is no noticeable outstanding feature of month 21 in the plots of residuals of `Patrons` in figure 2.

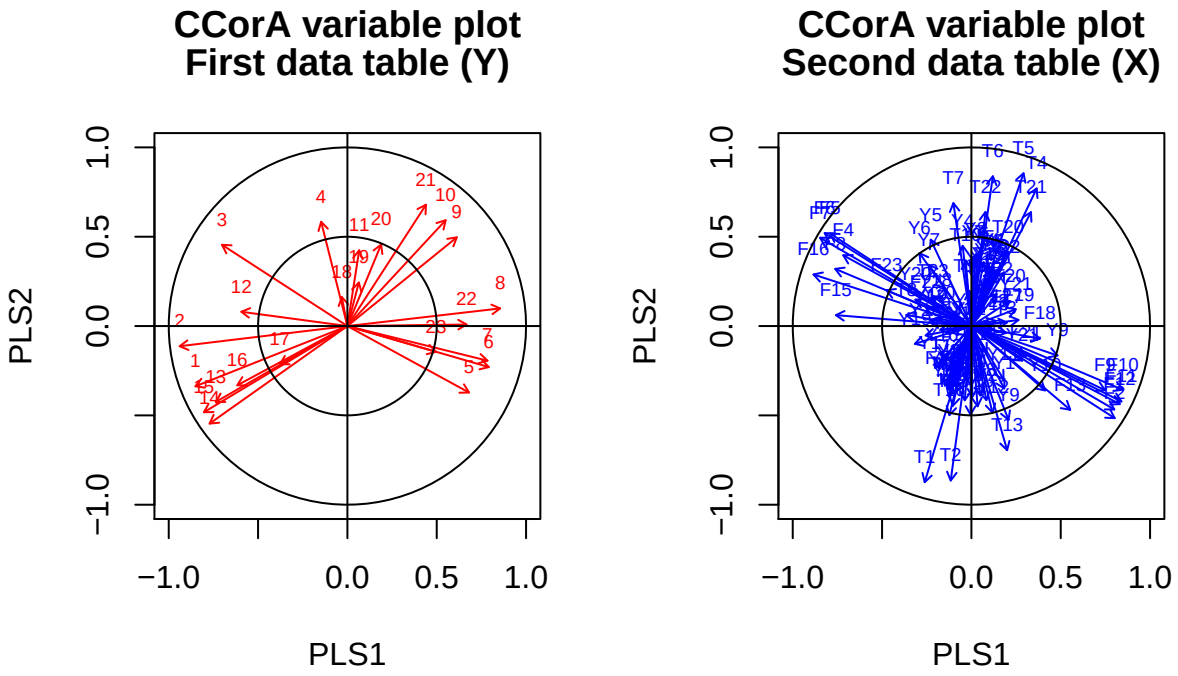


Figure 5: Biplot of Variables: plots the correlation of the variables (columns) with the first two PLS axes. On the left is the plot for Y variables while on the right is that for X variables.

3.4 X variables well represented by PLS

The biplots of variables for **X** (right plot in figure 5) indicate:

- Facebook Page Likes (F1 - F23 in the biplots) are well respresented by PLS axis 1
- Twitter Followers (T1 - T23 in the biplots) are well represented by PLS axis 2

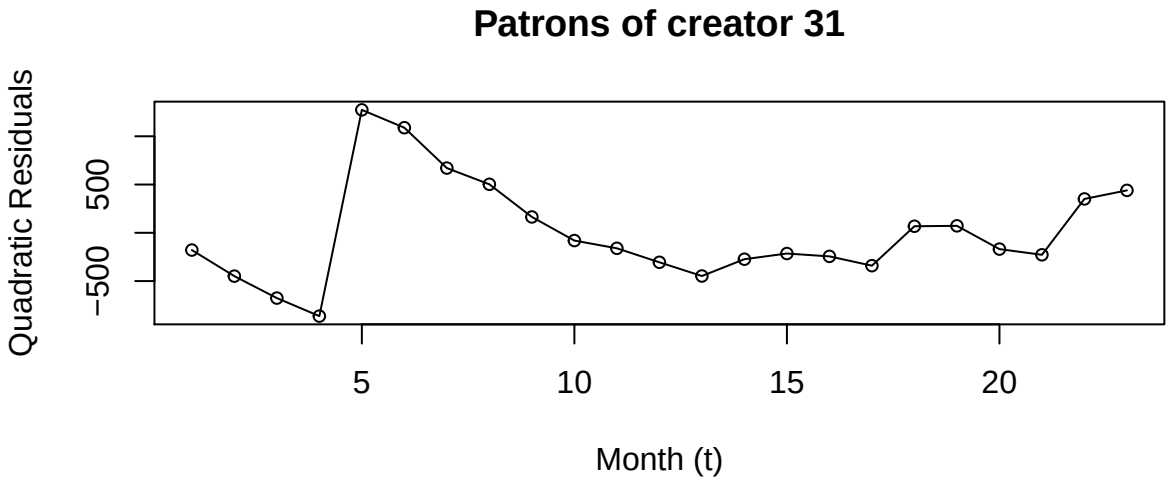
This implies that their relationship is strongest with variability in `Patrons`. The values are printed below for clarity.

3.5 Outliers

The biplots of objects in figure 4 indicate:

- For **Y** variables, there are atleast 2 outliers - creator 41 and 31. Figure 6 indicates their variability was very high during the period due to some large spikes. It implies that quadratic trend may not have been a good fit for them. Rather a shift in trend should have been modeled.
- For **X** variables, creators 41 and 31 are common outliers. Creator 13 is the third outlier. Figure 7 indicates their variability of Facebook.Likes was very high during the period due to some large spikes.

The trend for these creators may not have a followed a patter similar to others. The outliers may adversely affect our results and interpretation such as reducing correlations for other creators.



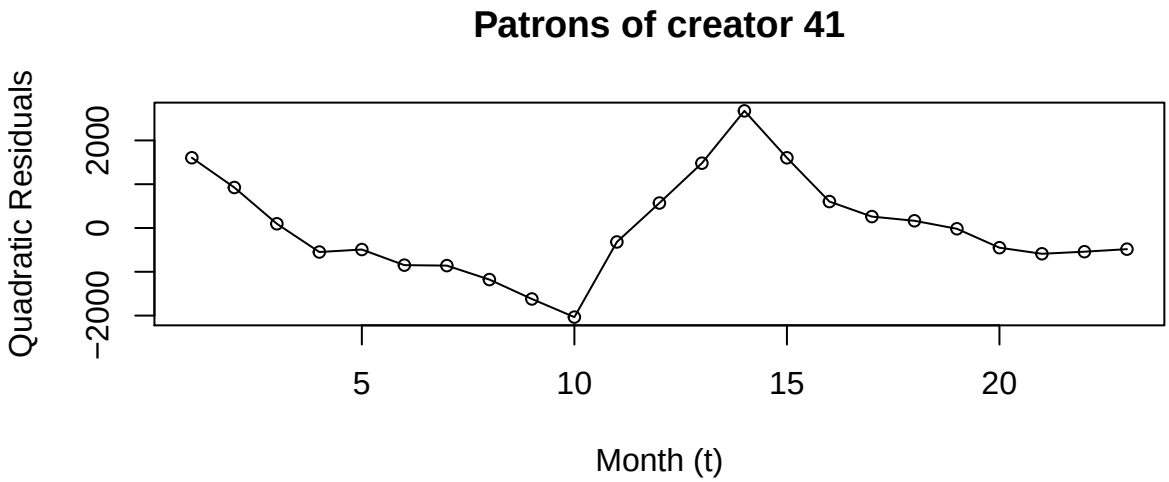


Figure 6: Quadratic Residuals of Patrons for outliers in Y data table

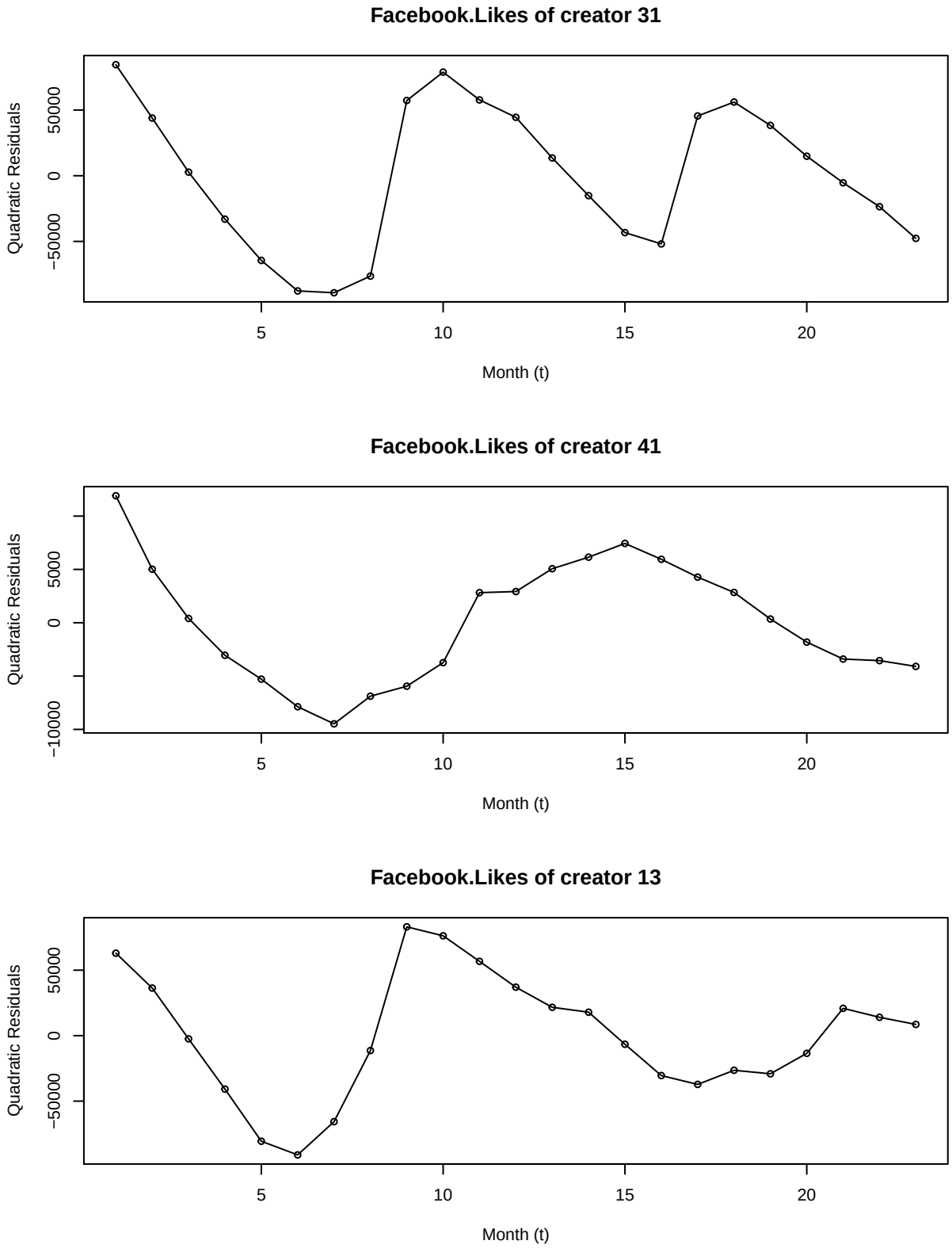


Figure 7: Quadratic Residuals of Facebook.Likes for outliers in X data table

3.6 PLS axis 1: Highest correlations with X

	Positive Corr.
Facebook.Likes_10	0.85
Facebook.Likes_11	0.84
Facebook.Likes_12	0.83
Facebook.Likes_2	0.80
Facebook.Likes_1	0.80
Facebook.Likes_9	0.75

	Negative Corr.
Facebook.Likes_16	-0.88
Facebook.Likes_7	-0.85
Facebook.Likes_6	-0.82
Facebook.Likes_5	-0.79
Facebook.Likes_8	-0.76
Facebook.Likes_15	-0.76

The first PLS axis for **X** has the highest correlations with `Facebook.Likes`.

- Months 5 to 8 have high negative correlation. This implies if `Facebook.Likes` don't increase, the score will be higher. For the corresponding first axis of Y, it indicated low variability.
- Months 9 to 12 have high positive correlation. They are not represented as clearly by axis 1 of Y and difficult to interpret. It could be a reason why the p-value was borderline.

3.7 PLS axis 2: Highest correlations with X

	Positive Corr.
Twitter.Followers_5	0.86
Twitter.Followers_6	0.84
Twitter.Followers_4	0.77
Twitter.Followers_7	0.69
Twitter.Followers_22	0.64
Twitter.Followers_21	0.64

	Negative Corr.
Twitter.Followers_1	-0.87
Twitter.Followers_2	-0.86
Twitter.Followers_13	-0.69
Youtube.Views_9	-0.52
Facebook.Likes_2	-0.52
Twitter.Followers_16	-0.50

The highest correlation of the second PLS axis is with `Twitter.Followers`.

- Months 4 to 7 and 21, 22 are positively correlated. The corresponding axis of Y is positively correlated

with months 9, 10 and 21. It could imply a time lag in the relationship between Twitter and `Patrons`.
- Negative correlations are weaker (closer to 0.5) and more random.
- It could be another reason for a borderline p-value.

3.8 PLS axes: Lowest absolute correlations with social metrics

PLS axis 1

	Abs. Corr.
Youtube.Views_14	0.00
Youtube.Videos_16	0.00
Twitter.Followers_14	0.01
Youtube.Views_1	0.01
Youtube.Videos_21	0.01
Youtube.Views_7	0.01

PLS axis 2

	Abs. Corr.
Youtube.Subscribers_23	0.00
Youtube.Videos_16	0.00
Youtube.Subscribers_13	0.02
Youtube.Subscribers_17	0.02
Facebook.Likes_17	0.02
Twitter.Followers_9	0.02

3.9 Lowest absolute correlations

YouTube dominates variables least correlated with the first two PLS axes.

- Videos are expected to have maximum viewer impact. `Youtube.Videos` may be considered to be a proxy for number of publications. In this context, it is a bit counter-intuitive.
- However, given that about half the creators aren't video publishers, it seems reasonable.
- Secondly, the ability to post videos on Facebook and Twitter seem to have taken their toll on YouTube.

4 Conclusions

Overall, there is a weak correlation between variability in social metrics and `Patrons`. Twitter and Facebook have higher correlation with `Patrons`. It reinforces popular belief that Twitter and Facebook are good platforms for engagement. However, it can have a negative impact as well. The time lag in the effect can be studied further. Effect of You Tube can be investigated separately for Video Creators.

The results may have been adversely affected to an extent as some of the creators had spikes and/or shifts in trend. Quadratic detrending may not have been the best way to detrend them.

4.1 What more can be done

Variables for most of the creators have an almost linear trend. If we regress them linearly over time, $t = (1, 2, \ldots, 23)$, we can get the slope and intercept term for each of the variables. PLS regression can be performed with slope and intercept terms of each metric as variables. This will let us know the effect of social media on the trend in `Patrons` rather than the variability.

5 Appendix: Code

```
knitr::opts_chunk$set(echo = FALSE, cache = TRUE, message=FALSE, warning = FALSE)
library(knitr)
library(mixOmics)
library(vegan)
set.seed(767)

load("data/saurabh767data2.rda")

include_graphics("figures/Patrons.pdf")

include_graphics("figures/Patrons_detrend.pdf")

par(mfrow = c(2,2))
qqnorm(MV_patreon[,2], main = "Patrons_1")
qqline(MV_patreon[,2])

qqnorm(MV_patreon[,3], main = "Facebook.Likes_1")
qqline(MV_patreon[,3])

qqnorm(MV_patreon[,4], main = "Twitter.Followers_1")
qqline(MV_patreon[,4])

qqnorm(MV_patreon[,5], main = "Youtube.Subscribers_1")
qqline(MV_patreon[,5])

# columns containing Patrons
Patrons.index <- grep("^Patron", colnames(MV_patreon) )

ytvideos.index <- grep("^Youtube.Videos", colnames(MV_patreon) )

fb.index <- grep("^Facebook", colnames(MV_patreon) )

category.index <- grep("^Category", colnames(MV_patreon) )

names.index <- grep("^Name", colnames(MV_patreon) )

# Response
Y <- as.matrix(MV_patreon[,Patrons.index ])
cat("Dimensions of Y\n")
dim(Y)

#cat("Names of first 10 Response Variables\n")
kable(colnames(Y)[1:10], col.names = "Names of first 10 Y Variables")


# Covariates
X <- as.matrix(MV_patreon[, -c(Patrons.index, category.index, names.index) ])
cat("Dimensions of X\n")
dim(X)

kable(colnames(X)[1:10], col.names = "Names of first 10 X variables")
```

```r
# PLS
all.pls <- pls(X, Y, mode="canonical")
observed <- cor(all.pls$variates[[1]][,1], all.pls$variates[[2]][,1])

socialcors <- rep(NA, 1000)

for(i in 1:1000){
perm <- sample(1:103)

mod <- pls(X, Y[perm,])

socialcors[i] <- cor(mod$variates[[1]][,1],mod$variates[[2]][,1])
}


# correlation
cat(paste0("Observed Correlation: " , round(observed, 2), "\n"))

#pvalue
cat(paste0("p-value of correlation based on 1000 permutations: " ,
           sum(socialcors > observed)/1000, "\n"))

observed2 <- cor(all.pls$variates[[1]][,2], all.pls$variates[[2]][,2])

socialcors2 <- rep(NA, 1000)

for(i in 1:1000){
perm <- sample(1:103)

mod <- pls(X, Y[perm,])

socialcors2[i] <- cor(mod$variates[[1]][,1],mod$variates[[2]][,1])
}

# correlation
cat(paste0("Observed Correlation: " , round(observed2, 2), "\n"))

#pvalue
cat(paste0("p-value of correlation based on 1000 permutations: " ,
           sum(socialcors2 > observed2)/1000, "\n"))
n <- ncol(Y)
patrons.cor <- matrix(NA, nrow= n, ncol = 2)

# correlation between original Y data and first two pls scores of Y
for(i in 1: n){
patrons.cor[i,1] <- cor(Y[,i], all.pls$variates$Y[,1])
patrons.cor[i,2] <- cor(Y[,i], all.pls$variates$Y[,2])
}

# correlation between covariates and first two pls scores of X

n <- ncol(X)
social.cor <- matrix(NA, nrow = n, ncol = 2)
```

```r
for(i in 1:n){
social.cor[i,1] <- cor(X[,i], all.pls$variates$X[,1])
social.cor[i,2] <- cor(X[,i], all.pls$variates$X[,2])
}

# name the rows
rownames(social.cor) <- colnames(X)
rownames(patrons.cor) <- 1:nrow(patrons.cor)

# column names
colnames(social.cor)<-c("PLS1", "PLS2")
colnames(patrons.cor)<-c("PLS1", "PLS2")

# relabel names for biplot

rnames <- rownames(social.cor)

first <- substr(rnames, start =1, stop = 1)

middle <- sub("^.*\\.([A-Z]).*" , "\\1" , rnames)

last <- sub("^[A-z]*.{1}[A-z]*_{1}([0-9]*)([0-9]+)$" , "\\1\\2" , rnames)

newnames <-paste0(first, last)

bi.social <- social.cor

rownames(bi.social) <- newnames



# create biplot object
social.bip <- list(Eigenvalues=rep(1,2), p.perm=NULL,
                   Cy = all.pls$variates[[2]],
                   Cx = all.pls$variates[[1]],
                   corr.Y.Cy = patrons.cor,
                   corr.X.Cx = bi.social)

class(social.bip) <- "CCorA"




# biplot

biplot(social.bip, plot.type= "objects" , cex = c(0.6, 0.6))


# biplot

biplot(social.bip, plot.type= "variables" , cex = c(0.6, 0.6))
```

```r
par(mfrow = c(2,1))

for(i in c(31, 41)){

creator31 <- unlist(MV_patreon[i, Patrons.index])
plot(creator31, main = paste0("Patrons of creator ", i),
     xlab = "Month (t)", ylab = "Quadratic Residuals",
     cex = 0.8)
lines(creator31)

}



par(mfrow = c(3,1))

for(i in c(31, 41, 13)){

creator31 <- unlist(MV_patreon[i, fb.index])
plot(creator31, main = paste0("Facebook.Likes of creator ", i),
     xlab = "Month (t)", ylab = "Quadratic Residuals",
     cex = 0.8)
lines(creator31)

}



sortedpls1 <- sort(social.cor[,1], decreasing = TRUE)

kable(round(head(sortedpls1), 2), col.names = "Positive Corr.")


sortedpls1 <- sort(social.cor[,1], decreasing = FALSE)

kable(round(head(sortedpls1), 2), col.names = "Negative Corr.")

sortedpls2 <- sort(social.cor[,2], decreasing = TRUE)

kable(round(head(sortedpls2), 2), col.names = "Positive Corr.")

sortedpls2 <- sort(social.cor[,2], decreasing = FALSE)

kable(round(head(sortedpls2), 2), col.names = "Negative Corr.")

pls1_low <- sort(abs(sortedpls1), decreasing = FALSE)

kable(round(head(pls1_low), 2), col.names = "Abs. Corr.")


pls2_low <- sort(abs(sortedpls2), decreasing = FALSE)

kable(round(head(pls2_low), 2), col.names = "Abs. Corr.")
```