Netflix Multi Regression

Stephanie Yang jy2777

12/5/2017

```
setwd("~/Desktop/Homework/Statistical Methods/Project/datasets")
#transpose t()
starbucks.metrics <- t(read.csv("netflix_metrics2.csv", header=FALSE))

#labels columns
colnames(starbucks.metrics)<- c("Date","Likes (Total) FB","Comments (Total) FB","Shares (Total) FB","Rea

#removes duplicate row
starbucks.metrics1 <- starbucks.metrics[-1,]

##Cleaning the metrics sheet
#1) removes space in column titles
colnames(starbucks.metrics1) <- gsub(" ","",colnames(starbucks.metrics1))

#2) removes % symbol of column 10
starbucks.metrics1[,c(11,15,19,20,21,26,27,28,31,33,38,39,43,44,52,59,60,61,69,71,72,73)] <- as.numeric
starbucks.metrics1 <- as.data.frame(starbucks.metrics1)
class(starbucks.metrics1)
```

```
## [1] "data.frame"
```

```
#3) removes comma separator for thousands, except for date column which is type character not numeric
#gsub to replace "," with "", and then convert the string to numeric using as.numeric
starbucks.metrics1[,2:73] <- lapply(starbucks.metrics1[,2:73], function(x) as.numeric(gsub(",","", as.cl
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
##Transforms Monthly to Quarterly Data:
library("lubridate")
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
#creates a year and quarter column per row
starbucks.metrics1$Date <- ymd(starbucks.metrics1$Date)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2017c.1.0/
## zoneinfo/America/New_York'
```

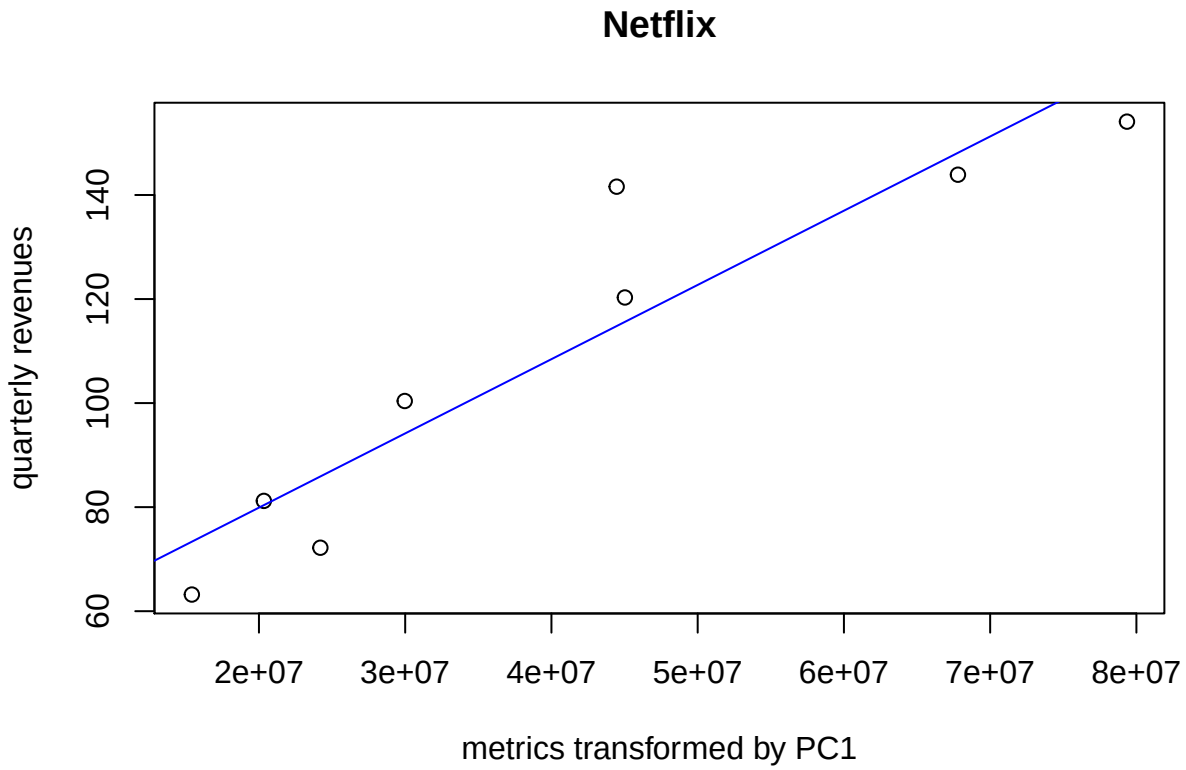
```
starbucks.metrics1$year = year(starbucks.metrics1$Date)
starbucks.metrics1$quarter = quarter(starbucks.metrics1$Date)

#aggregates quarters of same year and takes their sum (sales are aldo sums) : HOW TO DO IT WITH AGGREGA
library("reshape2")
starbucks.metrics2 <- melt(starbucks.metrics1[,2:75], id=c("quarter", "year"))
starbucks.metrics2 <- dcast(starbucks.metrics2, year + quarter ~ variable, fun.aggregate = sum)
write.csv(starbucks.metrics2,file="colgate_vizmetrics.csv")
```

```r
starbucks.metrics2 <- starbucks.metrics2[1:9,]

starbucks.metrics3 <- starbucks.metrics2[,-c(1,2)]
starbucks.metrics3 <- starbucks.metrics3[complete.cases(starbucks.metrics3),]

#************************************************************
#quarterly sales data, data points from CapitalIQ over 2 years
starbucks.sales <- read.csv("netflix_sales.csv")

#cleaning sales sheet: subsets,transposes and reformats data
starbucks.sales2 <- starbucks.sales[c(10,13),11:19]
starbucks.sales2 <- t(starbucks.sales2)
colnames(starbucks.sales2)<- c("Date","Sales")
starbucks.sales2[,2] <- as.numeric(gsub(",","",starbucks.sales2[,2]))
starbucks.sales2 <- as.data.frame(starbucks.sales2)
starbucks.sales2$Sales <- as.numeric(as.character(starbucks.sales2$Sales))

output.star <- prcomp(starbucks.metrics3)

PC1 <- output.star$rotation[,1]
PC2 <- output.star$rotation[,2]
PC3 <- output.star$rotation[,3]
PC4 <- output.star$rotation[,4]
PC5 <- output.star$rotation[,5]

PC1_data <- PC1%*%t(starbucks.metrics3)
PC2_data <- PC2%*%t(starbucks.metrics3)
PC3_data <- PC3%*%t(starbucks.metrics3)

netflix_sales <- starbucks.sales2$Sales[1:8]

metrics.transformed.by.PC1 <- as.numeric(PC1_data)
PC2_data <- as.numeric(PC2_data)
PC3_data <- as.numeric(PC3_data)

fit.net<- lm(netflix_sales~metrics.transformed.by.PC1)
summary(fit.net)
```

```
##
## Call:
## lm(formula = netflix_sales ~ metrics.transformed.by.PC1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.697 -10.244  -1.696   5.084  26.792
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.137e+01  1.072e+01   4.793 0.003022 **
## metrics.transformed.by.PC1 1.427e-06  2.323e-07   6.143 0.000852 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.13 on 6 degrees of freedom
```

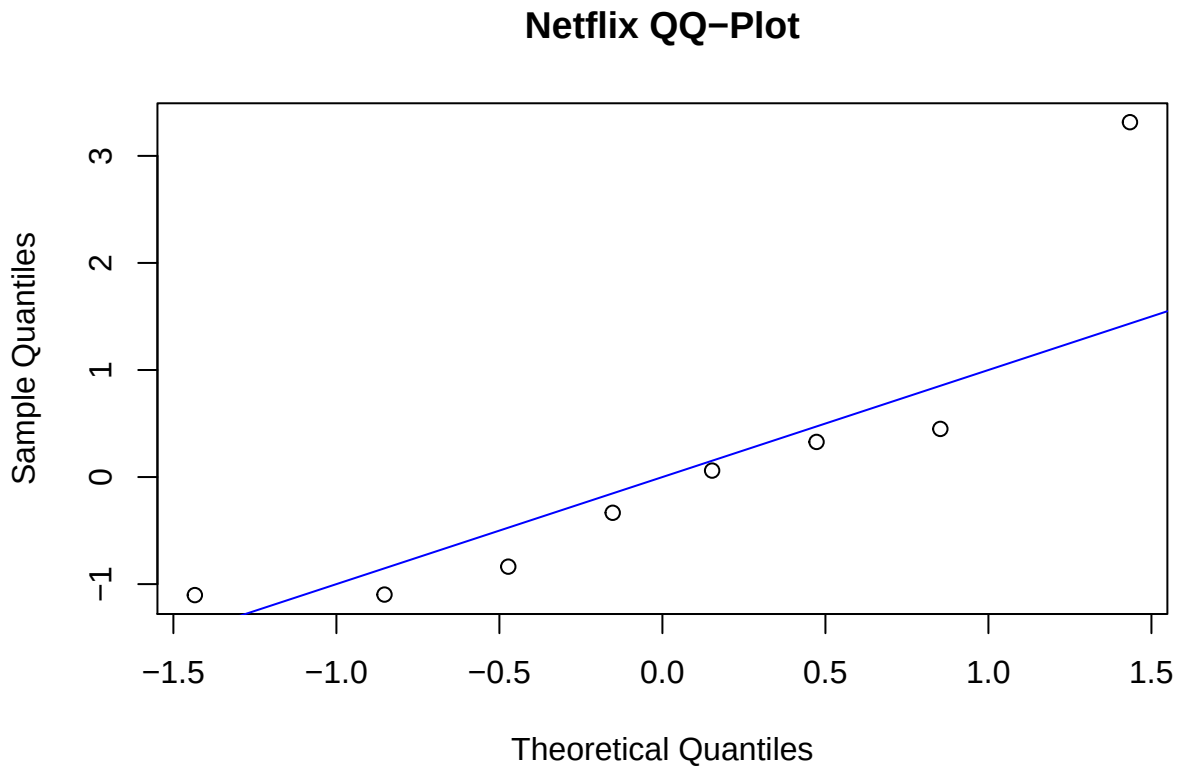
```
## Multiple R-squared:  0.8628, Adjusted R-squared:   0.84
## F-statistic: 37.74 on 1 and 6 DF,  p-value: 0.0008518
```

```r
# Linear regression plot
plot(metrics.transformed.by.PC1, netflix_sales, xlab = "metrics transformed by PC1", ylab = "quarterly r
abline(fit.net, col="blue")
```




```r
# Residual Plot
net.residual <- residuals(fit.net)
plot(net.residual, main = "Residual Plot for Netflix Regression")
abline(0,0, col="blue")
```


```
## QQ-plot
qqnorm(rstudent(fit.net), main = "Netflix QQ-Plot")
abline(0,1, col="blue")
```



```
library(arm)
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.9-3, built: 2016-11-21)
```

```
## Working directory is /Users/Stephanie/Desktop/Homework/Statistical Methods/Project/Slide codes
```

```
sim.1 <- sim(fit.net,1000)
meanPC <- mean(metrics.transformed.by.PC1)
Y.net <- sim.1@coef[,1]+sim.1@coef[,2]*meanPC+sim.1@sigma
hist(Y.net,seq(100,200,5))
```