

# Multivariate Regression and Monte Carlo

Stephanie Yang jy2777

12/3/2017

```
setwd("~/Desktop/Homework/Statistical Methods/Project/datasets")
#transpose t()
starbucks.metrics <- t(read.csv("starbucks_metrics2.csv", header=FALSE))

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

#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.c

## 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 also 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")
```

```

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("starbucks_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)
new_sales <- starbucks.sales2$Sales[1:8]

PC1_data <- as.numeric(PC1_data)

fit.star<- lm(new_sales~PC1_data)

summary(fit.star)

##
## Call:
## lm(formula = new_sales ~ PC1_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.61 -136.21  -38.57    1.38   443.12
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.774e+03  1.910e+02   9.286 8.82e-05 ***
## PC1_data     1.149e-05  4.210e-06   2.730  0.0342 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 211.7 on 6 degrees of freedom
## Multiple R-squared:  0.554, Adjusted R-squared:  0.4797
## F-statistic: 7.453 on 1 and 6 DF, p-value: 0.03419

```

```

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.star,1000)
meanPC <- mean(PC1_data)
Y.star <- sim.1@coef[,1]+sim.1@coef[,2]*meanPC+sim.1@sigma
hist(Y.star, seq(2000,3500,50))

```

**Histogram of Y.star**

