# 下载需要用到的包以及调用它们，没有安装的话用install.packages(“dplyr”)这样的格式安装即可

library(dplyr)

library(scales)

library(ggplot2)

# 读取文章用到的数据，我这里将数据改了名字

fund <- read.csv("C:/STONY/Practice/R (No.9)/fund.csv", header = T)

str(fund)

# 首先是数据处理阶段，先检查数据里是否有中途中止的众筹项目，这里只需要让每个项目的截至日期等于项目进行到最后一天就可以了，结果为没有中途中止的

### checking

f <- fund %>%

filter(max(as.Date(End\_Date,format='%m/%d/%Y'))== max(as.Date(Date,format='%m/%d/%Y')))

### there is no project which was stopped before the end date.

# 根据paper里处理数据的条件，得到数据子集。众筹项目大于5000，人数大于1，时间大于10.

### clean the data ###

### subset the data

fund1 <- fund %>%

filter(as.Date(End\_Date,format='%m/%d/%Y')==as.Date(Date,format='%m/%d/%Y'),Goal >= 5000,Backers >=1, Duration >= 10)

### 617 usable projects

# 计算出table2的数值，这里我分开计算了每一个指标的数值

### calculate the table 2

fund2 <- fund1 %>%

group\_by(Title,Category) %>%

summarise (goal = Goal,amount = Contributions,backers = Backers,

avg\_contri = Contributions/Backers,

per = Contributions/Goal\*100,num\_fri = Social\_Network\_Size,day = Duration)

# Campaign goal

summary(fund2$goal)

sd(fund2$goal)

# Amount raised

summary(fund2$amount)

sd(fund2$amount)

# Number of backers

summary(fund2$backers)

sd(fund2$backers)

# Average contribution per backer

summary(fund2$avg\_contri)

sd(fund2$avg\_contri)

# Percent of campaign goal raised

summary(fund2$per)

sd(fund2$per)

# Number of facebook friends

summary(fund2$num\_fri)

sd(fund2$num\_fri)

# Number of days

summary(fund2$day)

sd(fund2$day)

#计算出每个种类的项目数量，画出Fig2

### the distribution of campaigns across categories Fig2###

category <- fund1 %>%

group\_by(Category) %>%

summarise (n = n())

str(category)

#@@@@ how to change the " to others

ggplot(category,aes(Category,n,fill=Category)) +

geom\_bar(stat="identity",position="dodge") +

ggtitle("Category") +

scale\_y\_continuous(breaks=seq(0, 160, 20))

# 画出成功了以及没成功的项目的众筹款额变化曲线

# 有一点我有疑惑，图中说是nomalized的曲线，但是一般normalized的计算是（value-mean）/sd,很明显图里不是用的这个方法，因为图中横纵坐标就是直接计算得到的百分比值，所以这里因为paper没讲清楚，我就没做normalized.

### Normalized contribution dynamics of crowfunding campaigns Fig4 ###

fund3 <- subset(fund, ID = fund1$ID)

### calculate duartion % and progress %

fund3 <- fund3 %>%

group\_by(ID) %>%

mutate(dura = as.numeric((as.Date(Date,format='%m/%d/%Y')- as.Date(Start\_Date,format='%m/%d/%Y'))/max(Duration)\*100))

max(fund3$dura)

fund3 <- fund3 %>% group\_by(ID) %>% mutate(prog = Contributions/Goal\*100)

fund33 <- fund3

fund33 <- fund33 %>% filter(Funding\_Status <= 1)

# 这里我将数据分成了成功以及不成功，Funding\_Status=1成功，=0不成功

### In order to plot, we have to subset the data into success and unsuccess

fund330 <- fund33 %>% filter(Funding\_Status <= 0)

fund331 <- fund33 %>% filter(Funding\_Status == 1)

fund\_mean1 <- fund3 %>% filter(Funding\_Status == 1) %>% group\_by(dura) %>%

summarise(m\_prog=mean(prog))

fund\_mean0 <- fund3 %>% filter(Funding\_Status == 0) %>% group\_by(dura) %>%

summarise(m\_prog=mean(prog))

### 调用gridExtra，用于将几张图拼接在一张图里

library(gridExtra)

### @@@@@ smooth !!!!!!!!!

## P1画出数据里所有项目的中筹款目变化曲线，因为很多曲线的progress最大值已经远远超出150，一旦超出，ggplot就无法在ylim(0,150)中画出，所以此处，我做了一个subset,选取了在150以内的项目。但是很奇怪，做了subset后却无法显示曲线了，这个我和我同学讨论过，大家也没想出其他方法解决这个问题。**（1.这里后来我搜了很多代码，发现是span的设定不合理导致subset后的曲线也没法显示，后来我让span=1就可以画了。不好意思，我更新了代码，这里忘记更新了。2. 因为paper图里的最大值为150，所以需要将大于150的剔除。如果是你的自己的数据，可以限制在150或不限制，就根据你自己的要求来就可以了。也就是说fund33\_sub <- fund33 %>% filter(prog<=150) 这里的150你可以更改为你想要的数字，或者你跑自己的数据的时候，索性删除这一步）**

fund33\_sub <- fund33 %>% filter(prog<=150)

max(fund33\_sub$prog)

p1 <- ggplot() +

geom\_smooth(aes(x=dura, y=prog, group=ID, color=as.character(Funding\_Status)),se=FALSE,fund33\_sub) +

geom\_smooth(aes(x=dura, y=m\_prog, method="gam",colour = "red", lwd=1),se=FALSE,fund\_mean1) +

geom\_smooth(aes(x=dura, y=m\_prog, method="gam",colour = "red", lwd=1),fund\_mean0) +

ylim(c(0,150)) +

theme(legend.position="none")

ggplot() +

geom\_smooth(aes(x=dura, y=prog, group=ID, color=as.character(Funding\_Status)),se=FALSE,fund33) +

geom\_smooth(aes(x=dura, y=m\_prog, method="gam",colour = "red", lwd=1),se=FALSE,fund\_mean1) +

geom\_smooth(aes(x=dura, y=m\_prog, method="gam",colour = "red", lwd=1),fund\_mean0) +

ylim(c(-150,1500)) +

theme(legend.position="none")

## P2画出数据里成功项目的中筹款目变化曲线

## p2面临同样的问题

p2 <- ggplot() +

geom\_smooth(aes(x=dura, y=prog, group=ID),se=FALSE,fund331) +

geom\_smooth(aes(x=dura, y=m\_prog, method="gam",colour = "red", lwd=1),fund\_mean1) +

ylim(c(-150,1500)) +

theme(legend.position="none")

## P3画出数据里不成功项目的中筹款目变化曲线

p3 <- ggplot() +

geom\_smooth(aes(x=dura, y=prog, group=ID),se=FALSE,fund330) +

geom\_smooth(aes(x=dura, y=m\_prog, method="gam",colour = "red", lwd=1),fund\_mean0) +

ylim(c(0,150)) +

theme(legend.position="none")

grid.arrange(p1, p2, p3, nrow = 1)

### Fig5需要用到FPCA，由于paper里并没有提到它用的什么包，我就选用了R里较常用的FPCA的fdapace这个包。这里contribution trajectory同Fig4里的是一样的，velocity是FPCAfitted后的一阶数据，acceleration是二阶

#### for successful projects# cross-sectional

library(fdapace)

# 公式里的数不能为零，所以将为零的第2421行删除

fund331 <- fund331[-2412,]

# 将众筹项目的ID，时长%以及筹款进度%放进公式里进行计算

Files1 <- MakeFPCAInputs(fund331$ID, fund331$dura, fund331$prog)

fpcaObjFlies1 <- FPCA(Files1$Ly, Files1$Lt, list(plot=TRUE, methodMuCovEst='smooth', nRegGrid=25, FVEthreshold=1))

#以下为FPCA里成功项目中一阶和二阶的结果

fittedCurvesP11 <- fitted(fpcaObjFlies1, K=4, derOptns=list(p=1, kernelType="gauss"))

fittedCurvesP12 <- fitted(fpcaObjFlies1, K=4, derOptns=list(p=2, kernelType='gauss'))

# 同理，得到不成功的项目中一阶和二阶的结果

#### for unsuccessful projects

Files0 <- MakeFPCAInputs(fund330$ID, fund330$dura, fund330$prog)

fpcaObjFlies0 <- FPCA(Files0$Ly, Files0$Lt, list(plot=TRUE, methodMuCovEst='smooth'))

fittedCurvesP01 <- fitted(fpcaObjFlies0, derOptns=list(p=1, kernelType='epan'))

fittedCurvesP02 <- fitted(fpcaObjFlies0, derOptns=list(p=2, kernelType='epan'))

# 得到结果后，进行作图，注意！由于和paper里的用的包并不相同，所以不管是在smooth还是计算上都会有出入，但是不成功项目中的一阶图和paper里是相似的。这一点证明，虽然所用的包不相同，原理类似。

# 此处由于ggplot作图需要数据集是dataframe的格式，所以将结果放入新的数据集里便于作图，t\_temp 为时间，temp为一阶或二阶的值

### plot the velocity for success and unsuccess

t\_temp1 <- rep(fpcaObjFlies1$workGrid, length(Files1$Lid))

temp1 <- matrix(t(fittedCurvesP11), ncol=1)

id1 <- rep(1:length(Files1$Lid), each=25)

d1\_mat1 <- data.frame(cbind(id1, t\_temp1, temp1))

t\_temp0 <- rep(fpcaObjFlies0$workGrid, length(Files0$Lid))

temp0 <- matrix(t(fittedCurvesP01), ncol=1)

id0 <- rep(1:length(Files0$Lid), each=25)

d1\_mat0 <- data.frame(cbind(id0, t\_temp0, temp0))

der11 <- ggplot(d1\_mat1, aes(x=t\_temp1, y=temp1, group=id1)) +

geom\_line(aes(colour="red")) +

theme(legend.position="none")

der01 <- ggplot(d1\_mat0, aes(x=t\_temp0, y=temp0, group=id0)) +

geom\_line(aes(colour="red")) +

theme(legend.position="none")

grid.arrange(der11, der01, nrow = 1)

**（成功项目的值很大是因为有一些项目筹得的款额远远大于目标值，而不成功就没有这个问题，所以它的斜率就很小，波动不大，而成功项目中有斜率非常大的值，筹款速度极快。Paper里应该是删除了这样筹款速率极快的值，我没有，你仔细看会发现，除去成功项目中最大的那一条，整体的斜率值是在0-15间的。）**

### plot the velocity boxplot of success and unsuccess

dim(temp1)

dim(temp0)

temp1 <- cbind(temp1,rep(1,times=7300))

temp0 <- cbind(temp0,rep(0,times=16626))

velocity <- as.data.frame(rbind(temp1,temp0))

colnames(temp1) <- c("velocity","category")

ggplot(velocity, aes(category, velocity, fill=factor(category))) +

geom\_boxplot() +

ylim(c(0,3))

#### plot the accelaration of success and unsuccess

t\_temp1 <- rep(fpcaObjFlies1$workGrid, length(Files1$Lid))

temp1 <- matrix(t(fittedCurvesP12), ncol=1)

id1 <- rep(1:length(Files1$Lid), each=25)

d1\_mat1 <- data.frame(cbind(id1, t\_temp1, temp1))

t\_temp0 <- rep(fpcaObjFlies0$workGrid, length(Files0$Lid))

temp0 <- matrix(t(fittedCurvesP02), ncol=1)

id0 <- rep(1:length(Files0$Lid), each=25)

d1\_mat0 <- data.frame(cbind(id0, t\_temp0, temp0))

der12 <- ggplot(d1\_mat1, aes(x=t\_temp1, y=temp1, group=id1)) +

geom\_line(aes(colour="red")) +

theme(legend.position="none")

der02 <- ggplot(d1\_mat0, aes(x=t\_temp0, y=temp0, group=id0)) +

geom\_line(aes(colour="red")) +

theme(legend.position="none")

library(gridExtra)

grid.arrange(der12, der02, nrow = 1)

### plot the accelleration boxplot of success and unsuccess

dim(temp1)

dim(temp0)

temp1 <- cbind(temp1,rep(1,times=7300))

temp0 <- cbind(temp0,rep(0,times=16626))

accelleration <- as.data.frame(rbind(temp1,temp0))

colnames(accelleration) <- c("accelleration","category")

ggplot(accelleration, aes(category, accelleration, fill=factor(category))) +

geom\_boxplot() +

ylim(c(-0.02,0.01))

# 建立模型，将成功以及不成功的模型合并在一起，组建成新的我们建模需要用到的数据集。模型一共分为四种，Linear Regression, GAM,FDA,FDA+COVS. 此部分，针对不同的模型，根据progress的进度百分比依次进行建模。

### build prediction model

### get the new data set which contains velocity,accelaration, contractory, gaols.

fund\_331 <- fund331[-2412,]

fund\_model <- as.data.frame(rbind(fund\_331,fund330))

# 第一个是普通的线性回归模型

######################

#### LINEAR model ###

######################

# 建立mape公式，接下来都会用到，用以计算预测值和真实值之间的mape

### set mape function first, cause we are gonna use it later

mape <- function(actual,pred){

mape <- mean(abs((actual - pred)/actual))\*100

return (mape)

}

# 计算出10%数据下的每个项目的参与人数，筹得款额等，然后根据paper上的公式建立模型，通过summary()可得到 每个模型的R^2. 值得注意的是，paper选用了前518个项目进行建模，后10个进行预测，模型里都有按照paper的要求进行筛选

### 10 percent of model ###

LR\_dat10 <- fund\_model10 %>% group\_by(ID) %>% summarise(bak = max(Backers),

contri = max(Contributions),

soci = mean(Social\_Network\_Size),

goal = mean(Goal))

outcome10 <- fund\_model10 %>% group\_by(ID) %>% summarise(out = max(Contributions))

LR\_dat10 <- cbind(y = outcome10$out,

bak = LR\_dat10[,2],

contri = LR\_dat10[,3],

soci = LR\_dat10[,4],

goal = LR\_dat10[,5])

LR10 <- lm(y ~ bak+contri+goal+soci, data=as.data.frame(LR\_dat10[c(1:518),]))

summary(LR10)

AIC(LR10)

BIC(LR10)

# calculate the mape of LR10

# 模型预测，利用上述建立好的模型进行预测，利用519-618共10个项目的数据进行预测。得到预测值，而真实值就是10%里的每个项目的筹得款额，通过mape公式进行计算，得到mape值，以下的模型以此类推**（普通的线性回归中，R^2除了第一个10%，其他的都是随着数据集的增加而增加，而AIC/BIC也是一直在减小，只是它们的值都比较大）**

### predict 10 percent of campaign with 519-618 obs ###

new <- as.data.frame(LR\_dat10[c(519:618),])

pred <- predict(LR10, new, se.fit = TRUE)

actual <- as.data.frame(LR\_dat10[c(519:618),])$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

LR10.mape <- mape(data\_mape$y,data\_mape$yhat)

### 30 percent of model ###

LR\_dat30 <- fund\_model30 %>% group\_by(ID) %>% summarise(bak = max(Backers),

contri = max(Contributions),

soci = mean(Social\_Network\_Size),

goal = mean(Goal))

outcome <- fund\_model %>% group\_by(ID) %>% summarise(out = max(Contributions))

LR\_dat30 <- cbind(out = outcome[,2],

bak = LR\_dat30[,2],

contri = LR\_dat30[,3],

soci = LR\_dat30[,4],

goal = LR\_dat30[,5])

LR30 <- lm(out ~ bak+contri+goal+soci, data=as.data.frame(LR\_dat30[c(1:518),]))

summary(LR30)

AIC(LR30)

BIC(LR30)

## Adjusted R-squared: 0.9848

# calculate the mape of LR30

### predict 30 percent of campaign with 519-618 obs ###

new <- as.data.frame(LR\_dat30[c(519:618),])

pred <- predict(LR30, new, se.fit = TRUE)

actual <- as.data.frame(LR\_dat30[c(519:618),])$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

LR30.mape <- mape(data\_mape$y,data\_mape$yhat)

### 50 percent of model ###

LR\_dat50 <- fund\_model50 %>% group\_by(ID) %>% summarise(bak = max(Backers),

contri = max(Contributions),

soci = mean(Social\_Network\_Size),

goal = mean(Goal))

outcome <- fund\_model %>% group\_by(ID) %>% summarise(out = max(Contributions))

LR\_dat50 <- cbind(out = outcome[,2],

bak = LR\_dat50[,2],

contri = LR\_dat50[,3],

soci = LR\_dat50[,4],

goal = LR\_dat50[,5])

LR50 <- lm(out ~ bak+contri+goal+soci, data=as.data.frame(LR\_dat50[c(1:518),]))

summary(LR50)

AIC(LR50)

BIC(LR50)

## Adjusted R-squared: 0.9901

# calculate the mape of LR50

### predict 50 percent of campaign with 519-618 obs ###

new <- as.data.frame(LR\_dat30[c(519:618),])

pred <- predict(LR50, new, se.fit = TRUE)

actual <- as.data.frame(LR\_dat50[c(519:618),])$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

LR50.mape <- mape(data\_mape$y,data\_mape$yhat)

### 70 percent of model ###

LR\_dat70 <- fund\_model70 %>% group\_by(ID) %>% summarise(bak = max(Backers),

contri = max(Contributions),

soci = mean(Social\_Network\_Size),

goal = mean(Goal))

outcome <- fund\_model %>% group\_by(ID) %>% summarise(out = max(Contributions))

LR\_dat70 <- cbind(out = outcome[,2],

bak = LR\_dat70[,2],

contri = LR\_dat70[,3],

soci = LR\_dat70[,4],

goal = LR\_dat70[,5])

LR70 <- lm(out ~ bak+contri+goal+soci, data=as.data.frame(LR\_dat70[c(1:518),]))

summary(LR70)

AIC(LR70)

BIC(LR70)

## Adjusted R-squared: 0.9945

# calculate the mape of LR70

### predict 70 percent of campaign with 519-618 obs ###

new <- as.data.frame(LR\_dat70[c(519:618),])

pred <- predict(LR70, new, se.fit = TRUE)

actual <- as.data.frame(LR\_dat70[c(519:618),])$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

LR70.mape <- mape(data\_mape$y,data\_mape$yhat)

### 90 percent of model ###

LR\_dat90 <- fund\_model90 %>% group\_by(ID) %>% summarise(bak = max(Backers),

contri = max(Contributions),

soci = mean(Social\_Network\_Size),

goal = mean(Goal))

outcome <- fund\_model %>% group\_by(ID) %>% summarise(out = max(Contributions))

LR\_dat90 <- cbind(out = outcome[,2],

bak = LR\_dat90[,2],

contri = LR\_dat90[,3],

soci = LR\_dat90[,4],

goal = LR\_dat90[,5])

LR90 <- lm(out ~ bak+contri+goal+soci, data=as.data.frame(LR\_dat90[c(1:518),]))

summary(LR90)

AIC(LR90)

BIC(LR90)

## Adjusted R-squared: 0.9979

# calculate the mape of LR90

### predict 90 percent of campaign with 519-618 obs ###

new <- as.data.frame(LR\_dat90[c(519:618),])

pred <- predict(LR90, new, se.fit = TRUE)

actual <- as.data.frame(LR\_dat90[c(519:618),])$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

LR90.mape <- mape(data\_mape$y,data\_mape$yhat)

## 第二个模型是 FDA，也就是paper里选择的模型

#######################

#### FDA LINEAR Model #

#######################

# 同样，计算出模型中每个项目所需要的factor的值，然后按照paper里的公式建立模型，再利用mape公式计算mape **（关于FPCA的图的解释，第一个得看它说明书，但是grid的值可以看到你在FPCA() 里可以选择的最大指，例如图中grid最大为10，你就可以选择10以内的数值。我这里只看了怎么用。横着第二个应该是所有spline mean 关于时间的变化，这里看一个趋势就可以了，对于建模我目前没看到有什么影响。第三个是scree 图 指的是pca解释方差的比例，横坐标是主成分的个数，有几条断掉的横线就有几个。最后一个就是eigen function的图，它选取了主成分中的前三个成分而已，目前没看到对建模有什么影响。**

## 10 percent of campaign

library(fdapace)

fund\_model10 <- fund\_model %>% filter(dura <= 10)

Files\_model10 <- MakeFPCAInputs(fund\_model10$ID, fund\_model10$dura, fund\_model10$prog)

duration\_span <- 5

outcome10 <- sapply(Files\_model10$Ly, function(x) x[length(x)])

goal <- fund\_model10 %>% group\_by(ID) %>% summarise(goal=mean(Goal))

goal <- data.frame(goal[,2])

fpcaObjFlies\_model10 <- FPCA(Files\_model10 $Ly, Files\_model10 $Lt, list(plot=TRUE, methodMuCovEst='smooth', nRegGrid=10, FVEthreshold=1))

Contri\_pc\_score <- fpcaObjFlies\_model10$xiEst

fittedCurvesP\_m010 <- fitted(fpcaObjFlies\_model10, K=2, derOptns=list(p=0, kernelType="gauss"))

fittedCurvesP\_m110 <- fitted(fpcaObjFlies\_model10, K=2, derOptns=list(p=1, kernelType="gauss"))

fittedCurvesP\_m210 <- fitted(fpcaObjFlies\_model10, K=2, derOptns=list(p=2, kernelType='gauss'))

DM10 <- cbind( y = outcome10,

contri = fittedCurvesP\_m010[,1],

vel = fittedCurvesP\_m110[,1],

acce = fittedCurvesP\_m210[,1],

goal = goal)

FDA10 <- lm(y ~ contri+vel+goal, data=as.data.frame(DM10[c(1:518),]))

summary(FDA10)

AIC(FDA10)

BIC(FDA10)

# Adjusted R-squared: 0.9979

# calculate the mape of FDA10

### predict 10 percent of campaign with 519-618 obs ###

new <- DM10[c(519:618),]

pred <- predict(FDA10, new, se.fit = TRUE)

actual <- LR\_dat10[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDA10.mape <- mape(data\_mape$y,data\_mape$yhat)

## 30 percent of campaign

fund\_model30 <- fund\_model %>% filter(dura <= 30)

Files\_model30 <- MakeFPCAInputs(fund\_model30$ID, fund\_model30$dura, fund\_model30$prog)

duration\_span <- 5

outcome30 <- sapply(Files\_model30$Ly, function(x) x[length(x)])

goal <- fund\_model30 %>% group\_by(ID) %>% summarise(goal=mean(Goal))

goal <- data.frame(goal[,2])

fpcaObjFlies\_model30 <- FPCA(Files\_model30 $Ly, Files\_model30 $Lt, list(plot=TRUE, methodMuCovEst='smooth', nRegGrid=10, FVEthreshold=1))

Contri\_pc\_score30 <- fpcaObjFlies\_model30$xiEst

fittedCurvesP\_m030 <- fitted(fpcaObjFlies\_model30, K=2, derOptns=list(p=0, kernelType="gauss"))

fittedCurvesP\_m130 <- fitted(fpcaObjFlies\_model30, K=2, derOptns=list(p=1, kernelType="gauss"))

fittedCurvesP\_m230 <- fitted(fpcaObjFlies\_model30, K=2, derOptns=list(p=2, kernelType='gauss'))

DM30 <- cbind( y = outcome30,

contri = fittedCurvesP\_m030[,3],

vel = fittedCurvesP\_m130[,3],

acce = fittedCurvesP\_m230[,3],

goal = goal)

FDA30 <- lm(y ~ contri+vel+goal, data=as.data.frame(DM30[c(1:518),]))

summary(M30)

AIC(FDA30)

BIC(FDA30)

# Adjusted R-squared: 0.9997

# calculate the mape of FDA30

### predict 30 percent of campaign with 519-618 obs ###

new <- DM30[c(519:618),]

pred <- predict(FDA30, new, se.fit = TRUE)

actual <- LR\_dat30[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDA30.mape <- mape(data\_mape$y,data\_mape$yhat)

## 50 percent of campaign

fund\_model50 <- fund\_model %>% filter(dura <= 50)

Files\_model50 <- MakeFPCAInputs(fund\_model50$ID, fund\_model50$dura, fund\_model50$prog)

duration\_span <- 5

outcome50 <- sapply(Files\_model50$Ly, function(x) x[length(x)])

goal <- fund\_model50 %>% group\_by(ID) %>% summarise(goal=mean(Goal))

goal <- data.frame(goal[,2])

fpcaObjFlies\_model50 <- FPCA(Files\_model50 $Ly, Files\_model50 $Lt, list(plot=TRUE, methodMuCovEst='smooth', nRegGrid=10, FVEthreshold=1))

Contri\_pc\_score50 <- fpcaObjFlies\_model50$xiEst

fittedCurvesP\_m050 <- fitted(fpcaObjFlies\_model50, K=2, derOptns=list(p=0, kernelType="gauss"))

fittedCurvesP\_m150 <- fitted(fpcaObjFlies\_model50, K=2, derOptns=list(p=1, kernelType="gauss"))

fittedCurvesP\_m250 <- fitted(fpcaObjFlies\_model50, K=2, derOptns=list(p=2, kernelType='gauss'))

DM50 <- cbind( y = outcome50,

contri = fittedCurvesP\_m050[,5],

vel = fittedCurvesP\_m150[,5],

acce = fittedCurvesP\_m250[,5],

goal = goal)

FDA50 <- lm(y ~ contri+vel+goal, data=as.data.frame(DM50[c(1:518),]) )

summary(FDA50)

AIC(FDA50)

BIC(FDA50)

# Adjusted R-squared: 0.9987

# calculate the mape of FDA50

### predict 50 percent of campaign with 519-618 obs ###

new <- DM50[c(519:618),]

pred <- predict(FDA50, new, se.fit = TRUE)

actual <- LR\_dat50[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDA50.mape <- mape(data\_mape$y,data\_mape$yhat)

## 70 percent of campaign

fund\_model70 <- fund\_model %>% filter(dura <= 70)

Files\_model70 <- MakeFPCAInputs(fund\_model70$ID, fund\_model70$dura, fund\_model70$prog)

duration\_span <- 5

outcome70 <- sapply(Files\_model70$Ly, function(x) x[length(x)])

goal <- fund\_model70 %>% group\_by(ID) %>% summarise(goal=mean(Goal))

goal <- data.frame(goal[,2])

fpcaObjFlies\_model70 <- FPCA(Files\_model70 $Ly, Files\_model70 $Lt, list(plot=TRUE, methodMuCovEst='smooth', nRegGrid=10, FVEthreshold=1))

Contri\_pc\_score70 <- fpcaObjFlies\_model70$xiEst

fittedCurvesP\_m070 <- fitted(fpcaObjFlies\_model70, K=2, derOptns=list(p=0, kernelType="gauss"))

fittedCurvesP\_m170 <- fitted(fpcaObjFlies\_model70, K=2, derOptns=list(p=1, kernelType="gauss"))

fittedCurvesP\_m270<- fitted(fpcaObjFlies\_model70, K=2, derOptns=list(p=2, kernelType='gauss'))

DM70 <- cbind( y = outcome70,

contri = fittedCurvesP\_m070[,7],

vel = fittedCurvesP\_m170[,7],

acce = fittedCurvesP\_m270[,7],

goal = goal)

FDA70 <- lm(y ~ contri+vel+goal, data=as.data.frame(DM70[c(1:518),]) )

summary(FDA70)

AIC(FDA70)

BIC(FDA70)

#Adjusted R-squared: 0.9985

# calculate the mape of FDA70

### predict 70 percent of campaign with 519-618 obs ###

new <- DM70[c(519:618),]

pred <- predict(FDA70, new, se.fit = TRUE)

actual <- LR\_dat70[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDA70.mape <- mape(data\_mape$y,data\_mape$yhat)

## 90 percent of campaign

fund\_model90 <- fund\_model %>% filter(dura <= 90)

Files\_model90 <- MakeFPCAInputs(fund\_model90$ID, fund\_model90$dura, fund\_model90$prog)

duration\_span <- 5

outcome90 <- sapply(Files\_model90$Ly, function(x) x[length(x)])

goal <- fund\_model90 %>% group\_by(ID) %>% summarise(goal=mean(Goal))

goal <- data.frame(goal[,2])

fpcaObjFlies\_model90 <- FPCA(Files\_model90 $Ly, Files\_model90 $Lt, list(plot=TRUE, methodMuCovEst='smooth', nRegGrid=10, FVEthreshold=1))

Contri\_pc\_score90 <- fpcaObjFlies\_model90$xiEst

fittedCurvesP\_m090 <- fitted(fpcaObjFlies\_model90, K=2, derOptns=list(p=0, kernelType="gauss"))

fittedCurvesP\_m190 <- fitted(fpcaObjFlies\_model90, K=2, derOptns=list(p=1, kernelType="gauss"))

fittedCurvesP\_m290<- fitted(fpcaObjFlies\_model90, K=2, derOptns=list(p=2, kernelType='gauss'))

DM90 <- cbind( y = outcome90,

contri = fittedCurvesP\_m090[,9],

vel = fittedCurvesP\_m190[,9],

acce = fittedCurvesP\_m290[,9],

goal = goal)

FDA90 <- lm(y ~ contri+vel+goal, data=as.data.frame(DM90[c(1:518),]) )

summary(FDA90)

AIC(FDA90)

BIC(FDA90)

#Adjusted R-squared: 0.9981

# calculate the mape of FDA90

### predict 90 percent of campaign with 519-618 obs ###

new <- DM90[c(519:618),]

pred <- predict(FDA90, new, se.fit = TRUE)

actual <- LR\_dat90[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDA90.mape <- mape(data\_mape$y,data\_mape$yhat)

### this part can be ignored, just for checking

### full percent of campaign

Files\_model <- MakeFPCAInputs(fund\_model$ID, fund\_model$dura, fund\_model$prog)

duration\_span <- 5

outcome <- sapply(Files\_model$Ly, function(x) x[length(x)])

goal <- fund\_model %>% group\_by(ID) %>% summarise(goal=mean(Goal))

goal <- data.frame(goal[,2])

fpcaObjFlies\_model <- FPCA(Files\_model $Ly, Files\_model $Lt, list(plot=TRUE, methodMuCovEst='smooth', nRegGrid=10, FVEthreshold=1))

Contri\_pc\_score <- fpcaObjFlies\_model$xiEst

fittedCurvesP\_m0 <- fitted(fpcaObjFlies\_model, K=2, derOptns=list(p=0, kernelType="gauss"))

fittedCurvesP\_m1 <- fitted(fpcaObjFlies\_model, K=2, derOptns=list(p=1, kernelType="gauss"))

fittedCurvesP\_m2 <- fitted(fpcaObjFlies\_model, K=2, derOptns=list(p=2, kernelType='gauss'))

DM <- cbind( y = outcome,

contri = fittedCurvesP\_m0[,1],

vel = fittedCurvesP\_m1[,1],

acce = fittedCurvesP\_m2[,1],

goal = goal)

M <- lm(y ~ contri+vel+goal, data=as.data.frame(DM[c(1:518),]) )

summary(M)

# R^2 = 0.9981

AIC(M)

BIC(M)

# 第三个模型是加上了covariance的模型，paper里并没有讲加上的covs是什么，所以我就按照paper里所涉及到的变量，选择了soci为该模型中额外添加的变量

#######################

#### FDA+COV model ###

#######################

### the model with covariance

### 10 PERCENT MODEL ####

soci <- fund\_model10 %>% group\_by(ID) %>% summarise(soci=mean(Social\_Network\_Size))

soci <- data.frame(soci[,2])

DM10 <- cbind( y = outcome10,

contri = fittedCurvesP\_m010[,1],

vel = fittedCurvesP\_m110[,1],

acce = fittedCurvesP\_m210[,1],

goal = goal,

soci = soci)

FDACOV10 <- lm(y ~ contri+vel+goal+soci+acce, data=as.data.frame(DM10[c(1:518),]))

summary(M10)

BIC(FDACOV10)

AIC(FDACOV10)

# Adjusted R-squared: 0.9979

# calculate the mape of FDACOV10

### predict 10 percent of campaign with 519-618 obs ###

new <- DM10[c(519:618),]

pred <- predict(FDACOV10, new, se.fit = TRUE)

actual <- LR\_dat10[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDACOV10.mape <- mape(data\_mape$y,data\_mape$yhat)

### 30 PERCENT MODEL ####

soci <- fund\_mode30 %>% group\_by(ID) %>% summarise(soci=mean(Social\_Network\_Size))

soci <- data.frame(soci[,2])

DM30 <- cbind( y = outcome30,

contri = fittedCurvesP\_m030[,1],

vel = fittedCurvesP\_m130[,1],

acce = fittedCurvesP\_m230[,1],

goal = goal,

soci = soci)

FDACOV30 <- lm(y ~ contri+vel+goal+soci+acce, data=as.data.frame(DM30[c(1:518),]))

summary(FDACOV30)

BIC(FDACOV30)

AIC(FDACOV30)

# Adjusted R-squared: 0.9869

# calculate the mape of FDACOV30

### predict 30 percent of campaign with 519-618 obs ###

new <- DM30[c(519:618),]

pred <- predict(FDACOV30, new, se.fit = TRUE)

actual <- LR\_dat30[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDACOV30.mape <- mape(data\_mape$y,data\_mape$yhat)

### 50 PERCENT MODEL ####

soci <- fund\_mode50 %>% group\_by(ID) %>% summarise(soci=mean(Social\_Network\_Size))

soci <- data.frame(soci[,2])

DM50 <- cbind( y = outcome50,

contri = fittedCurvesP\_m050[,1],

vel = fittedCurvesP\_m150[,1],

acce = fittedCurvesP\_m250[,1],

goal = goal,

soci = soci)

FDACOV50 <- lm(y ~ contri+vel+goal+soci+acce, data=as.data.frame(DM50[c(1:518),]))

summary(FDACOV50)

BIC(FDACOV50)

AIC(FDACOV50)

# Adjusted R-squared: 0.9987

# calculate the mape of FDACOV50

### predict 50 percent of campaign with 519-618 obs ###

new <- DM50[c(519:618),]

pred <- predict(FDACOV50, new, se.fit = TRUE)

actual <- LR\_dat50[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDACOV50.mape <- mape(data\_mape$y,data\_mape$yhat)

### 70 PERCENT MODEL ####

soci <- fund\_mode70 %>% group\_by(ID) %>% summarise(soci=mean(Social\_Network\_Size))

soci <- data.frame(soci[,2])

DM70 <- cbind( y = outcome70,

contri = fittedCurvesP\_m070[,1],

vel = fittedCurvesP\_m170[,1],

acce = fittedCurvesP\_m270[,1],

goal = goal,

soci = soci)

FDACOV70 <- lm(y ~ contri+vel+goal+soci+acce, data=as.data.frame(DM70[c(1:518),]))

summary(FDACOV70)

BIC(FDACOV70)

AIC(FDACOV70)

# Adjusted R-squared: 0.999

# calculate the mape of FDACOV50

### predict 50 percent of campaign with 519-618 obs ###

new <- DM70[c(519:618),]

pred <- predict(FDACOV70, new, se.fit = TRUE)

actual <- LR\_dat70[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDACOV70.mape <- mape(data\_mape$y,data\_mape$yhat)

### 90 PERCENT MODEL ####

soci <- fund\_mode90 %>% group\_by(ID) %>% summarise(soci=mean(Social\_Network\_Size))

soci <- data.frame(soci[,2])

DM90 <- cbind( y = outcome90,

contri = fittedCurvesP\_m090[,1],

vel = fittedCurvesP\_m190[,1],

acce = fittedCurvesP\_m290[,1],

goal = goal,

soci = soci)

FDACOV90 <- lm(y ~ contri+vel+goal+soci+acce, data=as.data.frame(DM90[c(1:518),]))

summary(FDACOV90)

BIC(FDACOV90)

AIC(FDACOV90)

# Adjusted R-squared: 0.9981

# calculate the mape of FDACOV50

### predict 50 percent of campaign with 519-618 obs ###

new <- DM90[c(519:618),]

pred <- predict(FDACOV90, new, se.fit = TRUE)

actual <- LR\_dat90[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

FDACOV90.mape <- mape(data\_mape$y,data\_mape$yhat)

# 第四个模型，GAM模型，此处前期需要spline数据，paper里提到了用penalized的方法，所以这里用到了pspline这个包，计算NumCont，AmtCont，以它们spline之后的数值作为模型的变量。Paper中有提到。Mape的方法也同上述一样。

###################

#### GAM model ###

###################

# Penalized smooth spline

install.packages('pspline')

library(pspline)

##### 10 percent of progress for the GAM model

f\_NumCont\_tx\_10 <- rep(0, length(unique(fund\_model$ID)))

f\_AmtCont\_tx\_10 <- rep(0, length(unique(fund\_model$ID)))

marker <- 1

for(i in unique(fund\_model$ID)){

dat\_temp <- fund\_model[fund\_model$ID==i,]

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Backers, norder=2)

f\_NumCont\_tx\_10[marker] <- predict(spline\_temp, 10)

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Contributions, norder=2)

f\_AmtCont\_tx\_10[marker] <- predict(spline\_temp, 10)

marker <- marker + 1

}

f\_10 <- fund\_model10 %>% group\_by(ID) %>%

summarise(soci = max(Social\_Network\_Size),goal=max(Goal))

DM10 <- cbind( y = outcome10,

numcount = f\_NumCont\_tx\_10,

amtcount = f\_AmtCont\_tx\_10,

goal = f\_10[,3],

soci = f\_10[,2])

GAM10 <- lm(y ~ numcount+amtcount+goal+soci, data=as.data.frame(DM10[c(1:518),]))

summary(GAM10)

# Adjusted R-squared: 0.9666

AIC(GAM10)

BIC(GAM10)

# calculate the mape of GAM10

### predict 10 percent of campaign with 519-618 obs ###

new <- DM10[c(519:618),]

pred <- predict(GAM10, new, se.fit = TRUE)

actual <- LR\_dat10[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

GAM10.mape <- mape(data\_mape$y,data\_mape$yhat)

##### 30 percent of progress for the GAM model

f\_NumCont\_tx\_30 <- rep(0, length(unique(fund\_model$ID)))

f\_AmtCont\_tx\_30 <- rep(0, length(unique(fund\_model$ID)))

marker <- 1

for(i in unique(fund\_model$ID)){

dat\_temp <- fund\_model[fund\_model$ID==i,]

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Backers, norder=2)

f\_NumCont\_tx\_30[marker] <- predict(spline\_temp, 30)

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Contributions, norder=2)

f\_AmtCont\_tx\_30[marker] <- predict(spline\_temp, 30)

marker <- marker + 1

}

f\_30 <- fund\_model30 %>% group\_by(ID) %>%

summarise(soci = max(Social\_Network\_Size),goal=max(Goal))

DM30 <- cbind( y = outcome30,

numcount = f\_NumCont\_tx\_30,

amtcount = f\_AmtCont\_tx\_30,

goal = f\_30[,3],

soci = f\_30[,2])

GAM30 <- lm(y ~ numcount+amtcount+goal+soci, data=as.data.frame(DM30[c(1:518),]))

summary(GAM30)

# Adjusted R-squared: 0.9755

AIC(GAM30)

BIC(GAM30)

# calculate the mape of GAM30

### predict 30 percent of campaign with 519-618 obs ###

new <- DM30[c(519:618),]

pred <- predict(GAM30, new, se.fit = TRUE)

actual <- LR\_dat30[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

GAM30.mape <- mape(data\_mape$y,data\_mape$yhat)

##### 50 percent of progress for the GAM model

f\_NumCont\_tx\_50 <- rep(0, length(unique(fund\_model$ID)))

f\_AmtCont\_tx\_50 <- rep(0, length(unique(fund\_model$ID)))

marker <- 1

for(i in unique(fund\_model$ID)){

dat\_temp <- fund\_model[fund\_model$ID==i,]

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Backers, norder=2)

f\_NumCont\_tx\_50[marker] <- predict(spline\_temp, 50)

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Contributions, norder=2)

f\_AmtCont\_tx\_50[marker] <- predict(spline\_temp, 50)

marker <- marker + 1

}

f\_50 <- fund\_model50 %>% group\_by(ID) %>%

summarise(soci = max(Social\_Network\_Size),goal=max(Goal))

DM50 <- cbind( y = outcome50,

numcount = f\_NumCont\_tx\_50,

amtcount = f\_AmtCont\_tx\_50,

goal = f\_50[,3],

soci = f\_50[,2])

GAM50 <- lm(y ~ numcount+amtcount+goal+soci, data=as.data.frame(DM50[c(1:518),]))

summary(GAM50)

# Adjusted R-squared: 0.9649

AIC(GAM50)

BIC(GAM50)

# calculate the mape of GAM50

### predict 50 percent of campaign with 519-618 obs ###

new <- DM50[c(519:618),]

pred <- predict(GAM50, new, se.fit = TRUE)

actual <- LR\_dat50[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

GAM50.mape <- mape(data\_mape$y,data\_mape$yhat)

##### 70 percent of progress for the GAM model

f\_NumCont\_tx\_70 <- rep(0, length(unique(fund\_model$ID)))

f\_AmtCont\_tx\_70 <- rep(0, length(unique(fund\_model$ID)))

marker <- 1

for(i in unique(fund\_model$ID)){

dat\_temp <- fund\_model[fund\_model$ID==i,]

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Backers, norder=2)

f\_NumCont\_tx\_70[marker] <- predict(spline\_temp, 70)

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Contributions, norder=2)

f\_AmtCont\_tx\_70[marker] <- predict(spline\_temp, 70)

marker <- marker + 1

}

f\_70 <- fund\_model70 %>% group\_by(ID) %>%

summarise(soci = max(Social\_Network\_Size),goal=max(Goal))

DM70 <- cbind( y = outcome70,

numcount = f\_NumCont\_tx\_70,

amtcount = f\_AmtCont\_tx\_70,

goal = f\_70[,3],

soci = f\_70[,2])

GAM70 <- lm(y ~ numcount+amtcount+goal+soci, data=as.data.frame(DM70[c(1:518),]))

summary(GAM70)

# Adjusted R-squared: 0.9552

AIC(GAM70)

BIC(GAM70)

# calculate the mape of GAM70

### predict 70 percent of campaign with 519-618 obs ###

new <- DM70[c(519:618),]

pred <- predict(GAM70, new, se.fit = TRUE)

actual <- LR\_dat70[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

GAM70.mape <- mape(data\_mape$y,data\_mape$yhat)

##### 90 percent of progress for the GAM model

f\_NumCont\_tx\_90 <- rep(0, length(unique(fund\_model$ID)))

f\_AmtCont\_tx\_90 <- rep(0, length(unique(fund\_model$ID)))

marker <- 1

for(i in unique(fund\_model$ID)){

dat\_temp <- fund\_model[fund\_model$ID==i,]

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Backers, norder=2)

f\_NumCont\_tx\_90[marker] <- predict(spline\_temp, 90)

spline\_temp <- smooth.Pspline(dat\_temp$dura, dat\_temp$Contributions, norder=2)

f\_AmtCont\_tx\_90[marker] <- predict(spline\_temp, 90)

marker <- marker + 1

}

f\_90 <- fund\_model90 %>% group\_by(ID) %>%

summarise(soci = max(Social\_Network\_Size),goal=max(Goal))

DM90 <- cbind( y = outcome90,

numcount = f\_NumCont\_tx\_90,

amtcount = f\_AmtCont\_tx\_90,

goal = f\_90[,3],

soci = f\_90[,2])

GAM90 <- lm(y ~ numcount+amtcount+goal+soci, data=as.data.frame(DM90[c(1:518),]))

summary(GAM90)

# Adjusted R-squared: 0.9381

AIC(GAM90)

BIC(GAM90)

# calculate the mape of GAM90

### predict 90 percent of campaign with 519-618 obs ###

new <- DM90[c(519:618),]

pred <- predict(GAM90, new, se.fit = TRUE)

actual <- LR\_dat90[c(519:618),]$contri

data\_mape <- cbind(yhat=pred$fit,y=actual) %>% as.data.frame() %>% filter(pred$fit>0 & actual >0)

GAM90.mape <- mape(data\_mape$y,data\_mape$yhat)

# 模型建立完之后，我们便可以将R^2,BIC，AIC分别作图，由于paper里并没有提到spline以及FDA所调用的包是什么，所以我们的结果和paper有一些差异。并且paper里的数值，都是根据百分比的增加，BIC,AIC都随之变小，但我尝试的模型里的结果却不是这样的，有些甚至随之增大。我同我的师兄以及同学有探讨过这个问题，但是目前找不出原因，我们尝试建立了不同的模型，并且更换数据集，结果也还是如此。所以，此处不必太在意，但方法应是没有问题的。

### Fig 8 ####

### R Square ###

#得到每一个百分比阶段每种模型的R^2，并进行汇总，用于画图

R10 <- c(summary(LR10)["adj.r.squared"],summary(GAM10)["adj.r.squared"],

summary(FDA10)["adj.r.squared"],summary(FDACOV10)["adj.r.squared"])

R30 <- c(summary(LR30)["adj.r.squared"],summary(GAM30)["adj.r.squared"],

summary(FDA30)["adj.r.squared"],summary(FDACOV30)["adj.r.squared"])

R50 <- c(summary(LR50)["adj.r.squared"],summary(GAM50)["adj.r.squared"],

summary(FDA50)["adj.r.squared"],summary(FDACOV50)["adj.r.squared"])

R70 <- c(summary(LR70)["adj.r.squared"],summary(GAM70)["adj.r.squared"],

summary(FDA70)["adj.r.squared"],summary(FDACOV70)["adj.r.squared"])

R90 <- c(summary(LR90)["adj.r.squared"],summary(GAM90)["adj.r.squared"],

summary(FDA90)["adj.r.squared"],summary(FDACOV90)["adj.r.squared"])

Time <- c(10,30,50,70,90)

Rdata <- rbind(unlist(R10),unlist(R30),unlist(R50),unlist(R70),unlist(R90))

Rdata <- cbind(Rdata,unlist(Time))

colnames(Rdata) <- c("LR","GAM","FDA","FDACovs","Time")

Rdata <- as.data.frame(Rdata)

# 画图

ggplot(Rdata, aes(x = Time)) +

geom\_line(aes(y = LR), colour="blue") +

geom\_line(aes(y = GAM), colour = "grey") +

geom\_line(aes(y = FDA), colour = "red") +

geom\_line(aes(y = FDACovs), colour = "orange") +

ylab(label="Number of new members") +

xlab("Week")

### AIC and BIC ###

# 同上述的方式一样，汇总AIC和BIC的值

AIC10 <- c(AIC(LR10),AIC(GAM10),

AIC(FDA10),AIC(FDACOV10))

AIC30 <- c(AIC(LR30),AIC(GAM30),

AIC(FDA30),AIC(FDACOV30))

AIC50 <- c(AIC(LR50),AIC(GAM50),

AIC(FDA50),AIC(FDACOV50))

AIC70 <- c(AIC(LR70),AIC(GAM70),

AIC(FDA70),AIC(FDACOV70))

AIC90 <- c(AIC(LR90),AIC(GAM90),

AIC(FDA90),AIC(FDACOV90))

Time <- c(10,30,50,70,90)

AICdata <- rbind(unlist(AIC10),unlist(AIC30),unlist(AIC50),unlist(AIC70),unlist(AIC90))

AICdata <- cbind(AICdata,unlist(Time))

colnames(AICdata) <- c("LR","GAM","FDA","FDACovs","Time")

AICdata <- as.data.frame(AICdata)

ggplot(AICdata, aes(x = Time)) +

geom\_line(aes(y = LR), colour="blue") +

geom\_line(aes(y = GAM), colour = "grey") +

geom\_line(aes(y = FDA), colour = "red") +

geom\_line(aes(y = FDACovs), colour = "orange") +

ylab(label="Number of new members") +

xlab("Week")

BIC10 <- c(BIC(LR10),BIC(GAM10),

BIC(FDA10),BIC(FDACOV10))

BIC30 <- c(BIC(LR30),BIC(GAM30),

BIC(FDA30),BIC(FDACOV30))

BIC50 <- c(BIC(LR50),BIC(GAM50),

BIC(FDA50),BIC(FDACOV50))

BIC70 <- c(BIC(LR70),BIC(GAM70),

BIC(FDA70),BIC(FDACOV70))

BIC90 <- c(BIC(LR90),BIC(GAM90),

BIC(FDA90),BIC(FDACOV90))

Time <- c(10,30,50,70,90)

BICdata <- rbind(unlist(BIC10),unlist(BIC30),unlist(BIC50),unlist(BIC70),unlist(BIC90))

BICdata <- cbind(BICdata,unlist(Time))

colnames(BICdata) <- c("LR","GAM","FDA","FDACovs","Time")

BICdata <- as.data.frame(BICdata)

ggplot(BICdata, aes(x = Time)) +

geom\_line(aes(y = LR), colour="blue") +

geom\_line(aes(y = GAM), colour = "grey") +

geom\_line(aes(y = FDA), colour = "red") +

geom\_line(aes(y = FDACovs), colour = "orange") +

ylab(label="Number of new members") +

xlab("Week")

######################

#### PREDICTION ###

######################

# 此处计算出了mape的结果，将mape的结果汇总到数据框里，然后便于作图

### show the mape results and the plot

mape10 <- c(LR10.mape,GAM10.mape,

FDA10.mape,FDACOV10.mape)

mape30 <- c(LR30.mape,GAM30.mape,

FDA30.mape,FDACOV30.mape)

mape50 <- c(LR50.mape,GAM50.mape,

FDA50.mape,FDACOV50.mape)

mape70 <- c(LR70.mape,GAM70.mape,

FDA70.mape,FDACOV70.mape)

mape90 <- c(LR90.mape,GAM90.mape,

FDA90.mape,FDACOV90.mape)

Time <- c(10,30,50,70,90)

mapedata <- rbind(unlist(mape10),unlist(mape30),unlist(mape50),unlist(mape70),unlist(mape90))

mapedata <- cbind(mapedata,unlist(Time))

colnames(mapedata) <- c("LR","GAM","FDA","FDACovs","Time")

mapedata <- as.data.frame(mapedata)

# table3使用是整个数据集，然后采用了10-90的百分比进行数据集切割建模，table4采用了progress为30%以上的数据集，同样要进行切割建模，考虑到建模步骤仍然是上述4种model所用到的方法，内容相似且冗长，结果也与paper不同，此处我没有做30%以上的建模，如果你需要，我后续可以再做

### this is the table 3

head(mapedata)

### this is the fig9

ggplot(mapedata, aes(x = Time)) +

geom\_line(aes(y = LR), colour="blue") +

geom\_line(aes(y = GAM), colour = "black") +

geom\_line(aes(y = FDA), colour = "red") +

geom\_line(aes(y = FDACovs), colour = "orange") +

ylab(label="MAPE") +

xlab("Crowdfunding campaign time elapsed(%)")

# 画出fig10,按照paper给的条件，在20% duration 之内，progress在15-30之间的数据集

### this is the fig10

fund\_model\_uncer <- fund\_model %>%

filter(prog > 15 & prog < 30 & dura < 20)

fund\_model\_un <- subset(fund\_model,ID %in% fund\_model\_uncer$ID)

length(unique(fund\_model\_un$ID))

ggplot(fund\_model\_un,aes(x=dura, y=prog, group=ID, color=as.character(Funding\_Status))) +

geom\_smooth(span=1,se = FALSE) +

ylim(0,150) +

theme(legend.position="none")

PS： 这篇paper很多地方都讲得不清楚，例如数据spline，如何做的，使用的哪个包，FDA使用的自己写的包还是什么，都没讲明白，并且paper没有讲到在作图以及建模中也用到了其他的数据处理，比如数据集里progress百分比最大的已经上千了，但是它在呈现图的时候，仍然用的150的上限，这个时候应该需要数据处理。仔细读paper能够发现，它好像公式有一些地方也写错了。公式上一秒还在说outcomeT,但是MAPE下一秒计算的确是outcome T-X,真的很让人费解。我也把两种方式都尝试了一遍，最后和别人探讨总结说，模型公式写错了，应该是个typo。