Drivers of sample number

Engagement team - Freshawater Hackathon

25-27 May 2017

# Introduction

We want to know what the drivers of more samples is. We think this might be driven by things like the type of training they got, the amount and types of social engagement, length of involvement, etc.

# Data

The data were provided in .csv format. These were read in and some columns were fixed. The data was saved then as a .rdata which we read in here

## Warning: package 'lme4' was built under R version 3.3.3

## Loading required package: Matrix

## Warning: package 'ggplot2' was built under R version 3.3.3

## Warning: package 'reshape2' was built under R version 3.3.3

## Warning: package 'pscl' was built under R version 3.3.3

## Loading required package: MASS

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 3.3.3

## Classes and Methods for R developed in the

## Political Science Computational Laboratory

## Department of Political Science

## Stanford University

## Simon Jackman

## hurdle and zeroinfl functions by Achim Zeileis

## Warning: package 'glmm' was built under R version 3.3.3

## Loading required package: trust

## Warning: package 'trust' was built under R version 3.3.2

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.3.2

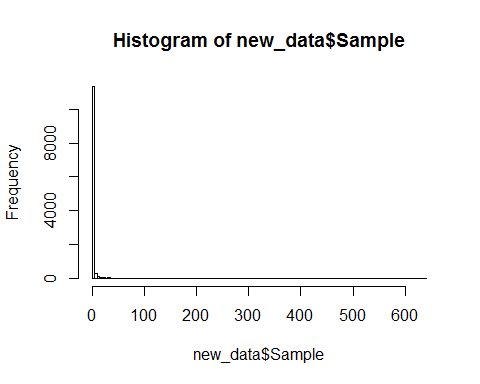
## Warning: package 'sjPlot' was built under R version 3.3.3

## Visit http://strengejacke.de/sjPlot for package-vignettes.

## Warning: package 'sjmisc' was built under R version 3.3.3

Looking at the data we need to look out for NAs and skew as well as the fact that our samples counts are zero inflated and skewed.

# We have a zero inflation problem  
hist(new\_data$Sample, breaks = 100)



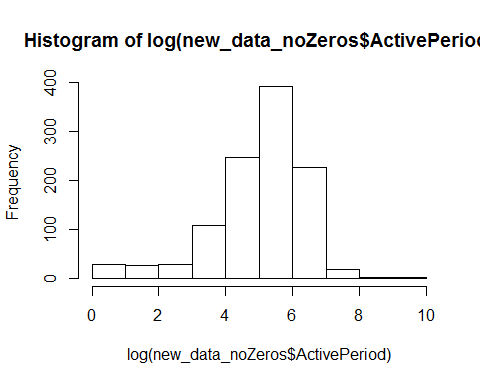
# get rid of zeros  
new\_data\_noZeros <- new\_data[new\_data$Sample > 0 & new\_data$ActivePeriod > 0 & new\_data$Staff == 0, ]  
  
# How much data do we have for the covariates of interest?  
nrow(new\_data\_noZeros)

## [1] 1323

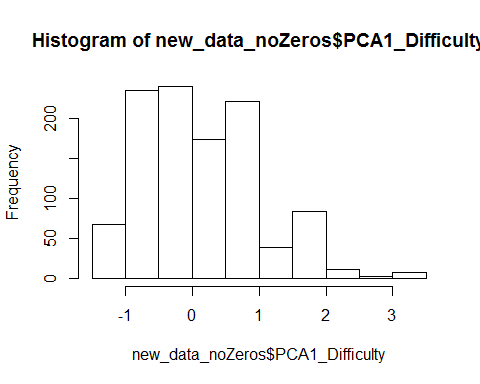
summary(new\_data\_noZeros[, c('ActivePeriod', 'Paid','Staff','PCA1\_Difficulty', 'WQS', 'Com\_score', 'Skill\_score', 'Team')])

## ActivePeriod Paid Staff PCA1\_Difficulty   
## Min. : 1.0 Min. :0.0000 Min. :0 Min. :-1.0279   
## 1st Qu.: 80.0 1st Qu.:0.0000 1st Qu.:0 1st Qu.:-0.7625   
## Median : 193.0 Median :0.0000 Median :0 Median :-0.1132   
## Mean : 298.6 Mean :0.4884 Mean :0 Mean : 0.1167   
## 3rd Qu.: 382.0 3rd Qu.:1.0000 3rd Qu.:0 3rd Qu.: 0.9710   
## Max. :10958.0 Max. :1.0000 Max. :0 Max. : 3.1795   
## NA's :246 NA's :678 NA's :246 NA's :246   
## WQS Com\_score Skill\_score Team   
## Min. :0 Min. : 0.000 Min. : 0.000 Min. :0.0000   
## 1st Qu.:0 1st Qu.: 0.000 1st Qu.: 2.000 1st Qu.:1.0000   
## Median :0 Median : 0.000 Median : 2.000 Median :1.0000   
## Mean :0 Mean : 7.892 Mean : 4.698 Mean :0.9023   
## 3rd Qu.:0 3rd Qu.: 0.000 3rd Qu.: 4.000 3rd Qu.:1.0000   
## Max. :0 Max. :2517.000 Max. :250.000 Max. :1.0000   
## NA's :246 NA's :246 NA's :246 NA's :678

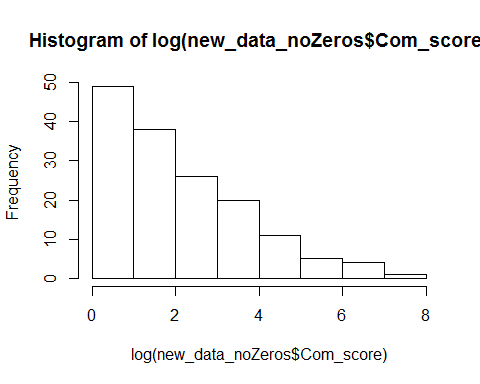
# We are going to lose lots of data through NAs in the Paid column  
# WQS would be great but there are so many NAs we just can't use it  
  
# We might be best loging some of our predictor variables  
hist(log(new\_data\_noZeros$ActivePeriod))



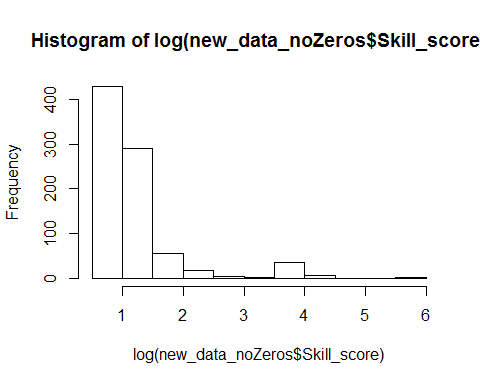
hist(new\_data\_noZeros$PCA1\_Difficulty)



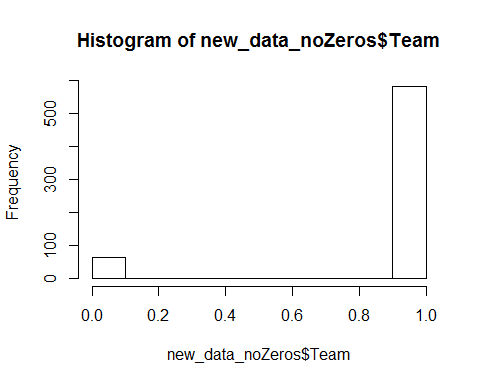
hist(log(new\_data\_noZeros$Com\_score))



hist(log(new\_data\_noZeros$Skill\_score))



hist(new\_data\_noZeros$Team)



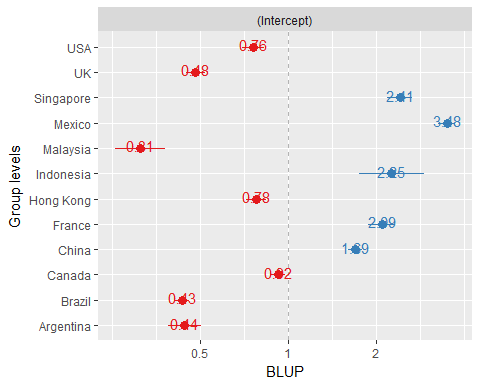
We can use a poisson distribution to account for the skew in teh samples counts and removes 0s to deal with the zero inflation. This step is reasonable as the 0 people are not involved in the project. The explanatory variables are in some cases improved by logging.

# We need to move to a mixed effects model because we want to put country in there  
m3 <- glmer(Sample ~ log(ActivePeriod+1) + Paid + PCA1\_Difficulty + log(Com\_score+1) + log(Skill\_score+1) + Team + (1|Country), data = new\_data\_noZeros, family = 'poisson')  
  
summary(m3)

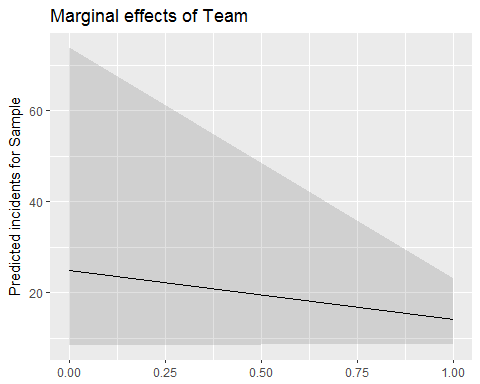
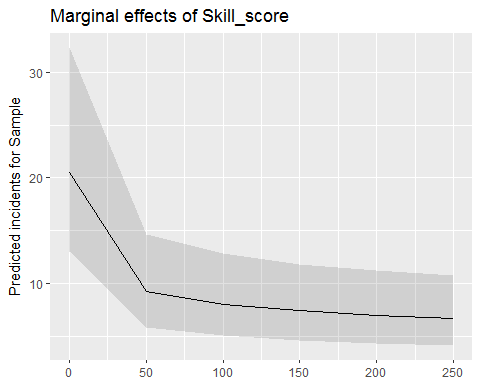
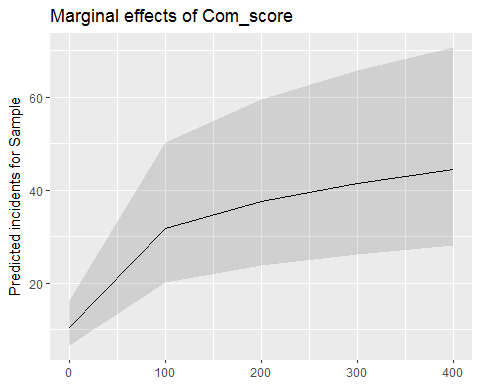
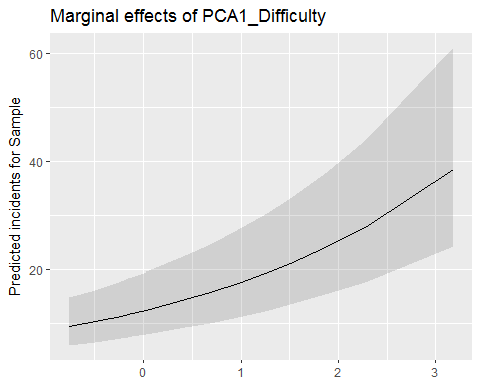
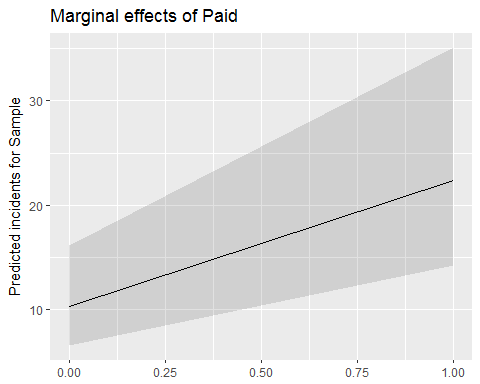
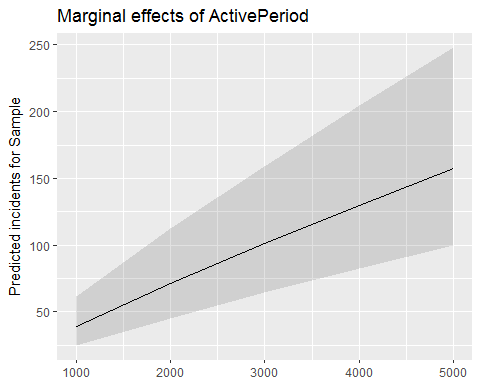
## Generalized linear mixed model fit by maximum likelihood (Laplace  
## Approximation) [glmerMod]  
## Family: poisson ( log )  
## Formula:   
## Sample ~ log(ActivePeriod + 1) + Paid + PCA1\_Difficulty + log(Com\_score +   
## 1) + log(Skill\_score + 1) + Team + (1 | Country)  
## Data: new\_data\_noZeros  
##   
## AIC BIC logLik deviance df.resid   
## 8137.1 8172.8 -4060.5 8121.1 637   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -8.790 -1.159 -0.096 1.092 40.720   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## Country (Intercept) 0.6063 0.7787   
## Number of obs: 645, groups: Country, 12  
##   
## Fixed effects:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.39364 0.56089 -4.27 1.98e-05 \*\*\*  
## log(ActivePeriod + 1) 0.86076 0.01451 59.31 < 2e-16 \*\*\*  
## Paid 0.77761 0.04273 18.20 < 2e-16 \*\*\*  
## PCA1\_Difficulty 0.35866 0.01997 17.96 < 2e-16 \*\*\*  
## log(Com\_score + 1) 0.24198 0.01038 23.30 < 2e-16 \*\*\*  
## log(Skill\_score + 1) -0.20187 0.01611 -12.53 < 2e-16 \*\*\*  
## Team -0.56510 0.60596 -0.93 0.351   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation of Fixed Effects:  
## (Intr) l(AP+1 Paid PCA1\_D l(C\_+1 l(S\_+1  
## lg(ActvP+1) -0.153   
## Paid -0.064 -0.138   
## PCA1\_Dffclt 0.049 0.025 -0.434   
## lg(Cm\_sc+1) 0.009 -0.209 0.227 0.010   
## lg(Skll\_+1) -0.037 0.059 -0.078 0.047 -0.277   
## Team -0.902 0.003 0.055 -0.060 0.011 0.000

# Adding country as a random effect has improved the model  
sjp.glmer(m3)

## Plotting random effects...



sjp.glmer(m3, type = "eff",  
 # axis.lim = list(c(0,50), c(0,5), c(0,5), c(0,5)),  
 facet.grid = FALSE,  
 show.p = TRUE,  
 show.ci = TRUE)



The results from the mixed effects model shows that a number of parameters are important. We plot the random effect which show how the countries vary in the amount of samples they generate and we plot the fixed effects. The increasing error in these i think reflects the fact that in most cases sample size decreases as values increase.