Science Forums

Cathy Su

19/10/2019

Executive Summary

Moderators at online forums are always interested in growing the participation while keeping a high quality discussion. Therefore, at times it becomes necessary to delete posts that might become problematic. The purpose of this study was to assess the variables that may influence which topics need to be closed on this online forum. It's been hypothesized that the type of post authors, topic of the post and XXX are factors in which discussions will need to be shut down. Discussion topics were randomly sampled from ScienceForums.Net (SFN). XXX studies passed the inclusion criteria and XXX variables represented potential contributing factors towards whether discussions will need to be closed. XXX proved to be the most consistent moderator of I-PA, suggesting that much of the discordance may be from motivational flux between initial intention and eventual behaviour. Anticipated regret and conscientiousness also had evidence as the moderators of I-PA. Perceived control/self-efficacy, planning, extraversion, habit and environmental proximity to recreation showed some evidence for moderation, while gender, agreeableness, openness, body mass index and ethnicity did not appear to moderate I-PA. The findings demonstrate that traditional intention theories may need augmentation to better account for the evidence present in I-PA discordance.

Introduction

Methods

Removal of outliers and missing data

Related to Figure 2, we removed the topics which are pinned since there seem to be very few cases so we may not be able to properly model what happened there. Looking at Figure 3, the only missing values in the data came from topics without categorization, which were also removed because we want to know the impact of the subforum type. Lastly we decided also on the basis of Figure 3 to remove topics with number of posts greater than 250 which seem to be extreme outliers.

Calculation of additional variables

For the substantive questions we calculated some new variables as follows:

- We converted the 'startdate' into POSIXct format which allows us to compare time elapsed precisely between dates. However to keep these numbers in a reasonable scale for the models, we subtracted 10^9 and divided the resulting number by the maximum startdate resulting in a (0,1] scale.
- the 'proportion_deleted' is the number of deleted divided by total posts in the topic which is necessary for Q1.
- The 'post_rate' is the number of posts divided by the 'duration' of the topic. If 'duration' is zero, the 'post_rate' was also assigned zero.
- 'posts_per_author' is number of posts divided by the number of distinct 'authors'.

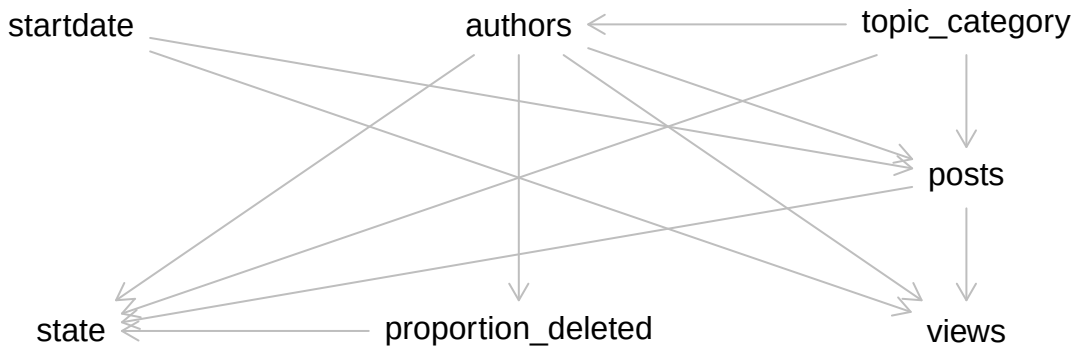

Figure 1: Causal diagram illustrates hypothesized relationships of experimental variables involved in relationship between proportion of deleted posts and topic status.

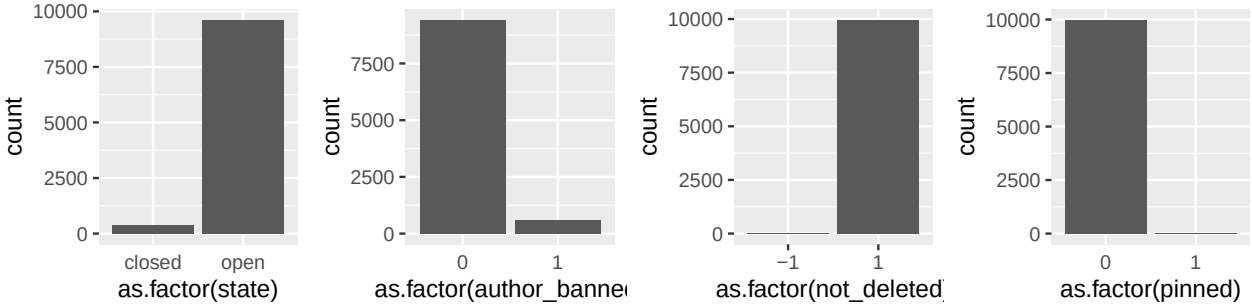

Figure 2: Distribution of binary categorical variables.

Exploratory Data Analysis

Causal diagram

Moderators would like to know which topics may need to be closed. We may hypothesize that topics will need to be closed mainly due to offensive posts, or due to controversial discussion. Based on this, author diversity ('authors') could be important to affect the relationship between proportion of deleted posts, length of the discussion and topic status (open or closed) as in Figure 1. This figure also shows that views on a topic

Univariate variable distributions

Figure 2 shows the distribution of topics which fall into each category for the binary variables. This shows us that almost no topics are deleted from view or pinned. However a small portion of topics (<10%) are closed, or started by a banned author.

Additionally, although topic type is an important variable of interest, we found that there were many topic ids which were missing a categorization (about 10%, see the top boxplot of Figure 3) which we removed. Additionally the number of posts was very right skewed, and we trimmed the few topics with posts in excess of 250 since these seem like extreme outliers based on Figure 3). The distribution of views was similarly right skewed to the distribution of posts. This suggests that it may be better to use a quasipoisson model than a poisson model for these counts. Indeed, when we tested for overdispersion with the package AER we found that the views and posts were both significantly overdispersed ($p < 0.001$, c> 5000 outputs not shown).

In Figure 4 we also see a relationship between duration, author_exp and startdate. If time increases by one unit in startdate, the number of views or posts may increase which suggests that we may want to use it as an offset.

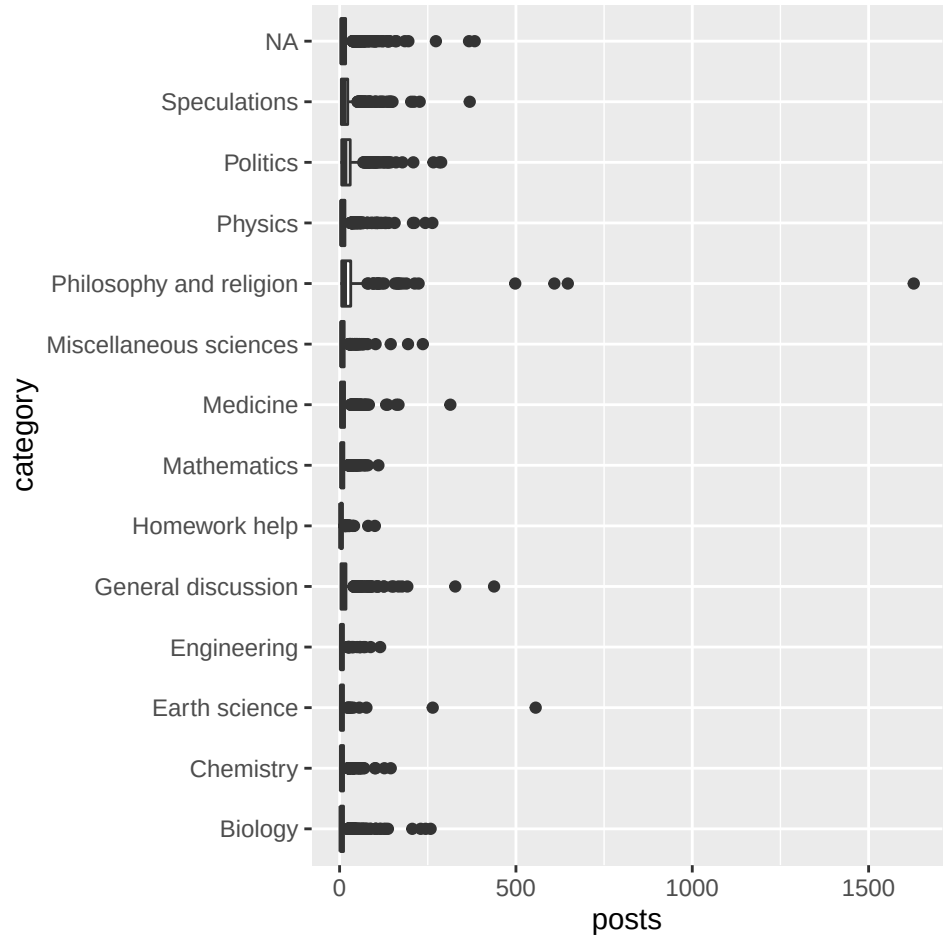

Figure 3: Boxplots giving breakdown of posts by subforum, showing many outlier topics which have an extreme number of posts.

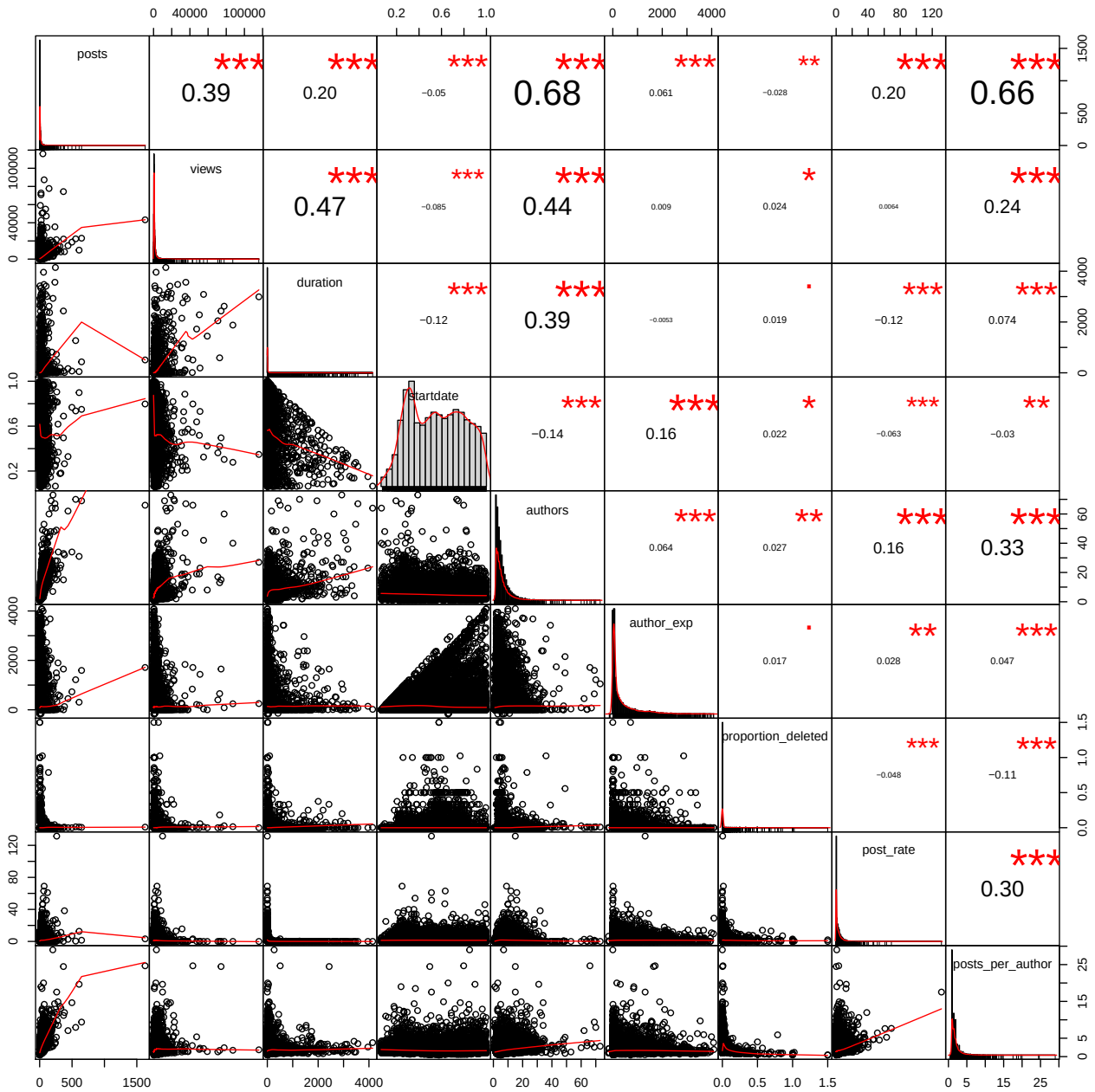

Figure 4: Pairwise correlations of important variables including their Pearson correlation coefficient. Significant correlations are marked by the corresponding number of red astericks. We can see from the univariate distributions (graphs on the diagonal) that with the exception of 'year_started', these variables are mostly very right skewed.

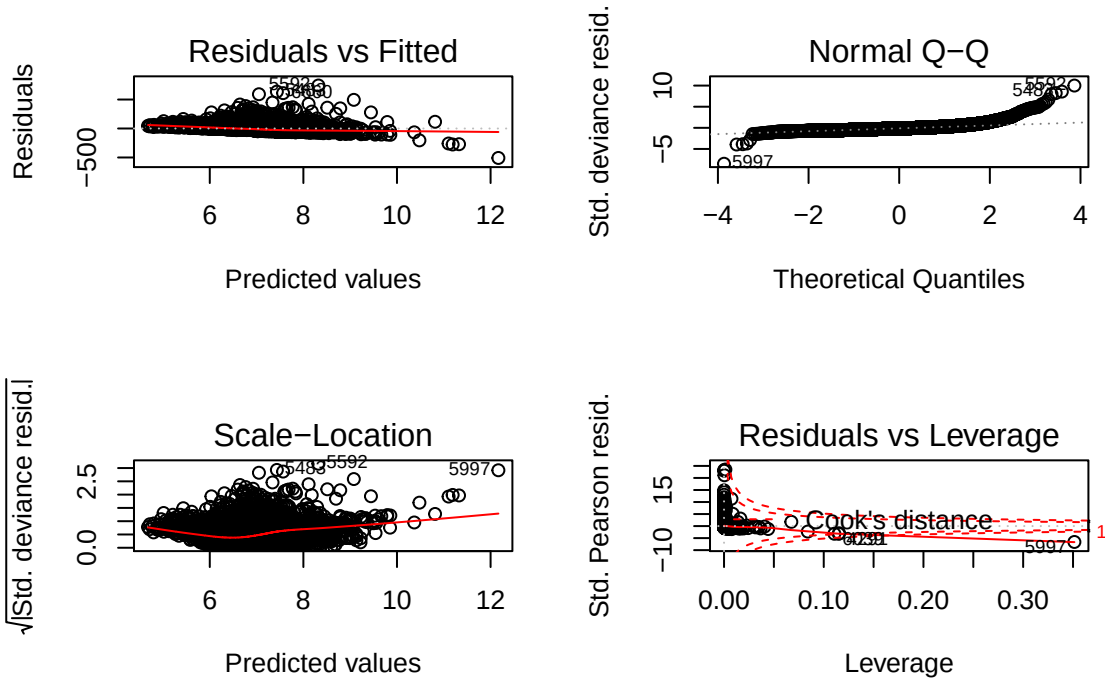

Figure 5: Diagnostic plots for the fit of model 'views ~ posts + authors + offset(log(startdate)'. Multiple outliers are apparent.

Q1. Relationship between views and posts

Since views are a form of counts which are positive and do not have a ceiling, to determine the relationship between views and posts, we start by fitting the glm with the quasipoisson family, log link and log('starttime') as offset. Based on our causal diagram, we chose to control for potential common cause 'authors' as well. The residuals of this basic model shown in Figure 5 suggest that the data are have multiple outliers. For example, the QQ plot shows that many residuals are not normally distributed, and these discussions e.g. row 5997, a 'Biology' topic, are especially prominent outliers in the residuals vs. leverage plot.

We then added a term for subforum. We compared the models of these with and without the additional variable 'category' (which relates to their subforum) by chi squared test, and the results are in Table 5. It seems that the relationship with views and posts varies strongly by subforum since the chi squared test suggests the latter model has a much better fit ($p < 0.001$).

```
##
##  Overdispersion test
##
## data:  mod1
## z = 7.459, p-value = 4.361e-14
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   5506.706
```

Table 1: Chi squared test models with and without controlling for subforum

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9018	17961982	NA	NA	NA

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9006	16482045	12	1479937	0

Q2. Diverse discussions closed or deleted

To check whether author diversity affects topic state, we modelled the responses 'state' and 'not_deleted' against the predictor 'authors'. Since the response is binary we used a quasibinomial family glm with logit link, with 'starttime' as offset. Based on our causal diagram we also controlled for subforum and number of posts. These were decent models of the topics that were open and not deleted but did not represent the closed and deleted topics very well (LHS of Figure **??**). However with authors added as covariate, the number of outliers is much less (RHS of Figure **??**). Furthermore the result of chi squared test is given in Table 2-3. This suggests that the model with authors better represents the data and discussions involving more authors are significantly more likely to be closed ($p < 0.05$) or deleted ($p < 0.001$).

Table 2: Comparison of model of state with and without controlling for authors

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9007	2485.352	NA	NA	NA
9006	2476.999	1	8.354	0.006

Table 3: Comparison of model of deleted topics with and without controlling for authors

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9007	240.594	NA	NA	NA
9018	273.393	-11	-32.799	0.002

Q3 Do topics with deleted posts tend to get closed more often?

To check whether deleted posts tend to get closed more often, we checked if the full model of the response 'state' from Q2 could be improved by adding the predictor 'proportion_deleted' by chi squared test. The inclusion of proportion deleted significantly improves the model ($p < 0.001$).

Table 4: Comparison of model of state from Q2 with and without controlling for proportion of deleted posts

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9006	2476.999	NA	NA	NA
9005	2475.439	1	1.56	0.233

Q4. Are members who have been registered for longer before starting the topic more

successful at starting active discussions?

To check whether long registered members may be more successful, we modelled the response 'posts' as proxy

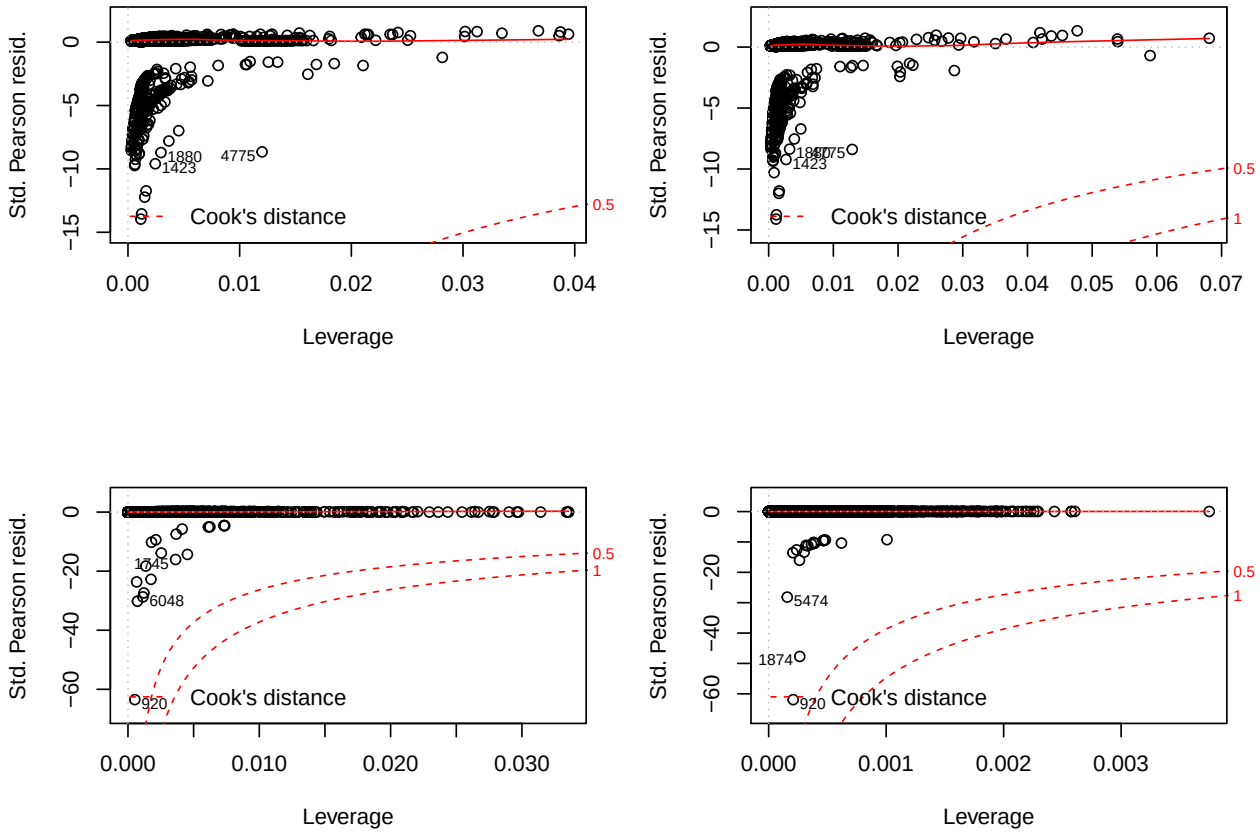

Figure 6: Residual plots of each of the models of 'state' (top row) and 'not_deleted' (bottom row). R

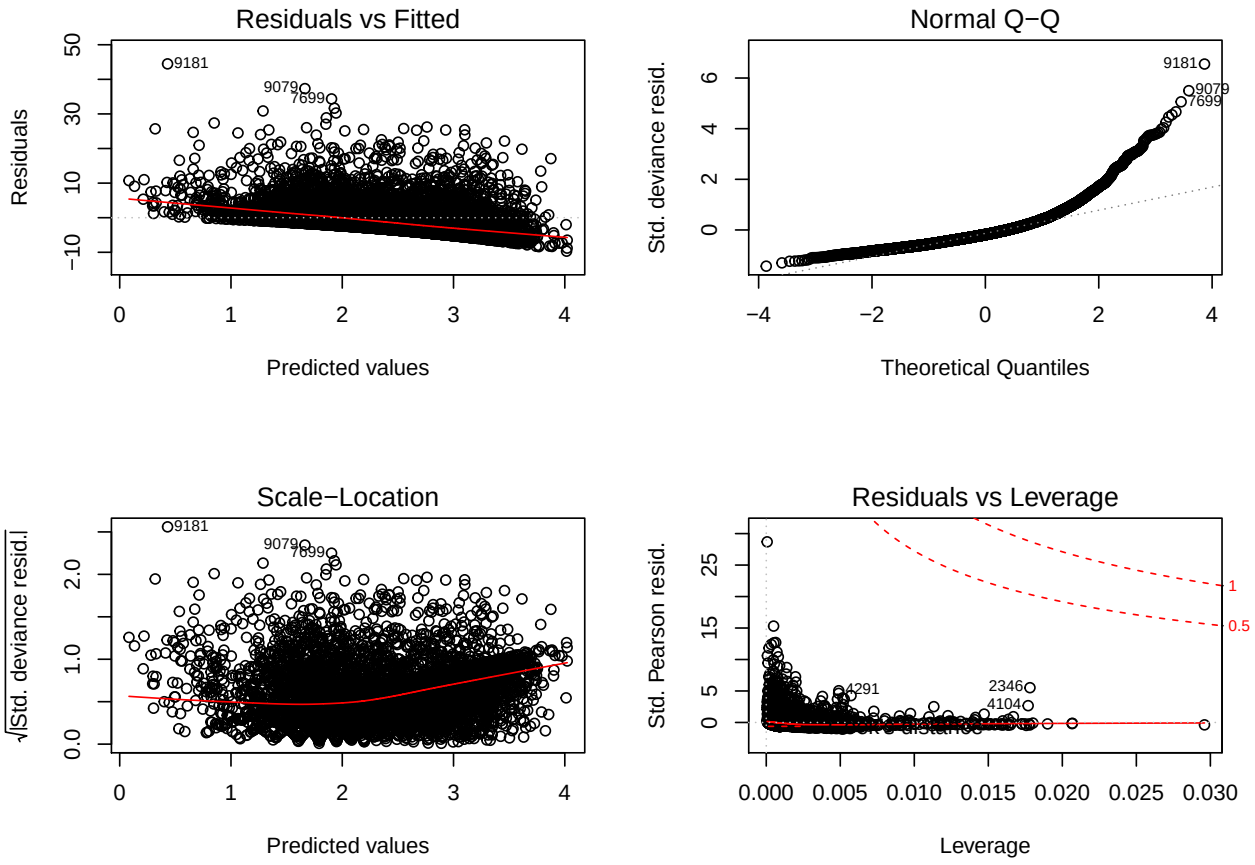

Figure 7: Diverse discussions

for activity of the topic against the predictor 'author_exp'. Since the response is a positive count we use quasipoisson with log link. From our causal diagram we decided to control for subforum with log(startdate) as offset.

Q5. Predicting whether a given topic will be closed

To build a classification model to predict whether a given topic will be closed, we chose 'state' as the response variable and then selected the potential predictors based upon whether they would be available while the topic is active. This means we could pick from only the following variables:

```
vars <- colnames(dat)[c(6,12, 13, 14:16)]
vars
```

```
## [1] "startdate"          "author_exp"         "author_banned"
## [4] "category"           "proportion_deleted" "post_rate"
```

To build a classification model we divided our data randomly into 5-fold and used 1 fold as the test set. First we picked the best glm model using the training set, we performed best subsets regression using exhaustive search and AIC with the 'bestglm' package. Since the response is a binary variable, we chose models from the binomial family with logit link. Due to the categorical variables, we used AIC instead of cross validation to pick the best model since some subsets will not contain all the categories. The coefficients of the best model is shown in Table **??**. We compared the performance of the model on the test set against the performance of the

```
# dat$y <- dat$state
# newvars <- c(vars, "y")
#
# testIndexes <- sample(nrow(dat),round(nrow(dat)/10))
# testData <- dat[testIndexes,newvars]
# trainData <- dat[-testIndexes,newvars]
#
# # k folds.
#   # best fit on the train data (regsubsets )
#   best.fit=bestglm(Xy=trainData,
#                    family = binomial,
#                    IC = "AIC",
#                    method="exhaustive")
#
#   best.fit$BestModels
#   res <- summary(best.fit$BestModel)
#   # performance on the test data
# #   for(i in 1:10){ #
# #     pred=predict(best.fit,team_dat[folds==k,vars],id=i)
# #     cv.errors[k,i]=mean((team_dat$final.performance[folds==k]-pred)^2)
# #   }
# # }
# #source("logit_dotplot.R")
# #logit_dotplot()
# x <- predict(best.fit$BestModel, newdata = testData)
#
#
# mean((testData$y-x)^2)
#   # PLOT PREDICTIONS
#   # https://www.barelysignificant.com/post/glm/
#   # d %>%
#   #   ggplot(aes(x = weight, y = height) ) +
#   #   geom_line(aes(y = predict(mod1)), size = 1) +
#   #   geom_point(size = 2, alpha = 0.3) +
#   #   geom_segment(
#   #       x = wght, xend = wght,
#   #       y = 0, yend = predict(mod1, newdata = data.frame(weight = wght) ),
#   #       linetype = 2) +
#   #   geom_segment(
#   #       x = 0, xend = wght,
#   #       y = predict(mod1, newdata = data.frame(weight = wght) ),
#   #       yend = predict(mod1, newdata = data.frame(weight = wght) ),
#   #       linetype = 2) +
#   #   theme_bw(base_size = 12)
#
#
# kable(coef(res, 8), format = "latex",caption = "Coefficients of model picked by best subsets regressi
```

Results

The results discussed by the original study (**???**) include that:

- considered in isolation, group diversity and testosterone are not significantly correlated with performance.

- when group diversity was l

Model selection

Since the authors studied 2-way interactions, to choose the terms in the model we first performed model selection using best subsets and cross validation. We start off with all of the group level variables as depicted in Figure 4, adding based on the causal graph and the authors' work the following interaction terms: