Replication of RQ5-6 regression analyses

Data preparation

We load the file with the scores from LIWC and rescale them in the range [1,5].

```
personality = read_delim(params$data, ";", escape_double = FALSE)
personality$openness <- resca(personality,openness,
                              new_min=1, new_max=5)$openness_res
personality$conscientiousness <- resca(personality,conscientiousness,
                                       new_min=1, new_max=5)$conscientiousness_res
personality$extraversion <- resca(personality,extraversion,
                                  new_min=1, new_max=5)$extraversion_res
personality$agreeableness <- resca(personality,agreeableness,
                                   new_min=1, new_max=5)$agreeableness_res
personality$neuroticism <- resca(personality,neuroticism,
                                 new_min=1, new_max=5)$neuroticism_res
```

Now we load the commit data and merge them with the personality data.

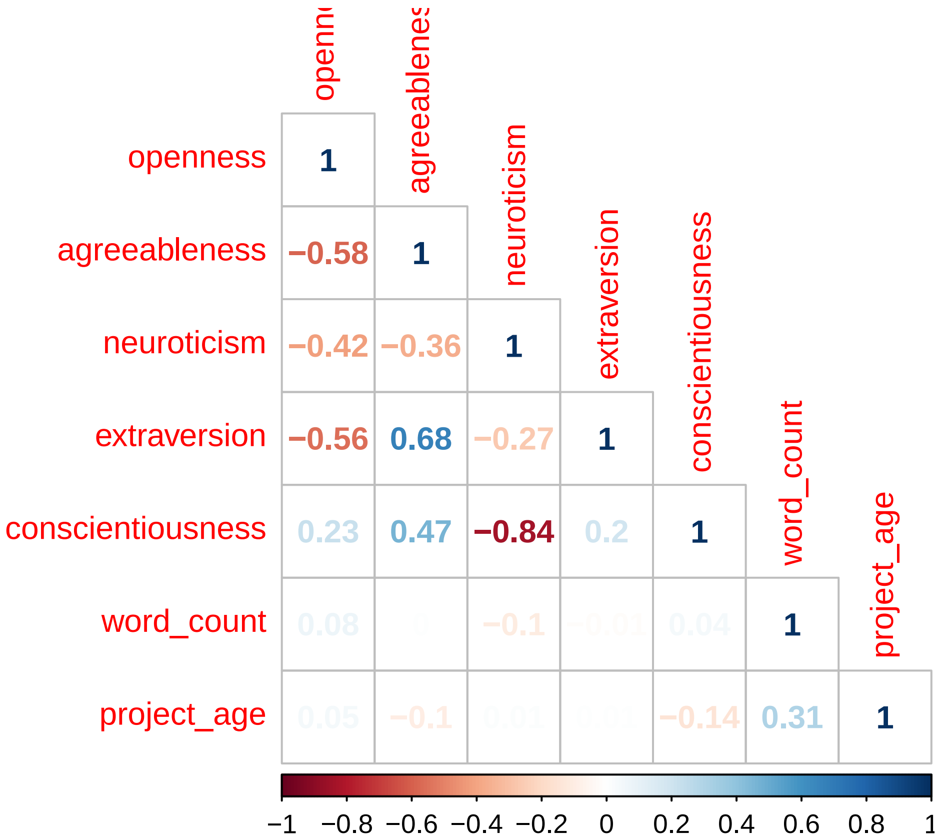
```
commit = read_delim(params$commits, ";", escape_double = FALSE)
# Find which developers appear in the intersection of the two data sets
both = intersect(unique(commit$uid), unique(personality$uid))
# Extract data only for the intersection developers
# (filter out people with 0 commits)
commit_both = subset(commit, uid %in% both &
                       num_authored_commits > 0)
id_no_cotributors = setdiff(unique(personality$uid), unique(commit$uid))
pers_zero_contributors = subset(personality, uid %in% id_no_cotributors)

commit.count = sqldf(
  "select uid, sum(num_authored_commits) as 'total_commits',
  count(project) as 'total_projects',
  max(last_authored_datetime) as last_authored_datetime,
  min(first_integrated_datetime) as first_integrated_datetime
  from `commit_both` group by uid"
)

# Filter out people who are still active (have at least one commit
# during the last 3 months before data collection). We can't know
# if they will remain one-time contributors or have more commits
commit.count = subset(commit.count,
                      last_authored_datetime < as.POSIXct("2017-09-01 20:18:02"))

# Identify people with only one commit total, across all projects.
# These are the one-timers. The others are more active, even if they
# have projects with only one commit
one.timers = subset(commit.count, total_commits == 1)$uid
multi.timers = subset(commit.count, total_commits > 1)$uid
oneormore.timers = subset(commit.count, total_commits >= 1)$uid
```

```
# Assign a binary label "one_timer" to everyone in the personality
# data, based on the distinction above
p = subset(personality, uid %in% one.timers | uid %in% multi.timers)
p$one_timer = FALSE
for (i in 1:nrow(p)) {
  p[i,]$one_timer = p[i,]$uid %in% one.timers
}

p = subset(personality, uid %in% one.timers | uid %in% multi.timers)

# Compute average personality scores per person, across time and across all projects
p.aggr = sqldf(
  "select uid, avg(openness) as 'openness',
  avg(agreeableness) as 'agreeableness',
  avg(neuroticism) as 'neuroticism',
  avg(extraversion) as 'extraversion',
  avg(conscientiousness) as 'conscientiousness',
  sum(word_count) as word_count,
  project
  from p group by uid"
)

# Apply the "one_timer" label to this aggregate data set
p.aggr$one_timer = FALSE
for (i in 1:nrow(p.aggr)) {
  p.aggr[i,]$one_timer = p.aggr[i,]$uid %in% one.timers
}
# Apply the "mulit_timer" label to this aggregate data set
p.aggr$m_timer = FALSE
for (i in 1:nrow(p.aggr)) {
  p.aggr[i,]$m_timer = p.aggr[i,]$uid %in% multi.timers
}
# calculate project age
proj_age = sqldf::sqldf("select project, project_age
                        from `commit`
                        group by project")
# add this pieace of info to the aggregated dataset
p.aggr = p.aggr[p.aggr$project %in% unique(proj_age$project),]
p.aggr$project_age = 0
for (i in 1:nrow(p.aggr)) {
  p.aggr[i,]$project_age =
    proj_age[proj_age$project == p.aggr[i,]$project,]$project_age
}
```

RQ5: logistic regression model of contribution likelihood

We build a simple logistic regression model to explain whether someone is a one-time contributor or not based on their personality score, controlling for the number of words in their emails (which may influence personality). Accordingly, the dependent, predicted variable is whether or not a one-time contributor will make further contributions.

First we analyze potential high correlations in the dataset. We drop *neuroticism* and *extraversion* because they shows high correlation (~70) with *conscientiousness* and *agreeableness*, respectively.



```
m = glm(
  m_timer ~ log(word_count) +
    project_age +
    openness +
    #agreeableness +
    #neuroticism +
    extraversion +
    conscientiousness,
  data = p.aggr,
  family = binomial(link = 'logit')
)
```

The results show that only the control variable *project age* has a significant, negative effect.

```
arm::display(m)
```

```
## glm(formula = m_timer ~ log(word_count) + project_age + openness +
##     extraversion + conscientiousness, family = binomial(link = "logit"),
##     data = p.aggr)
##                   coef.est coef.se
## (Intercept)        1.31    12.87
## log(word_count)   -0.06     0.15
## project_age       -0.37     0.08
## openness          -0.53     2.82
## extraversion       0.54     2.50
## conscientiousness  1.03     3.02
## ---
##   n = 106, k = 6
##   residual deviance = 84.5, null deviance = 115.8 (difference = 31.3)
```

```
summary(m)
```

```
##
## Call:
## glm(formula = m_timer ~ log(word_count) + project_age + openness +
##     extraversion + conscientiousness, family = binomial(link = "logit"),
##     data = p.aggr)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.288   0.338   0.460   0.515   1.724
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.3100    12.8665    0.10     0.92
## log(word_count)    -0.0582     0.1456   -0.40     0.69
## project_age        -0.3656     0.0804   -4.55  5.4e-06 ***
## openness           -0.5324     2.8226   -0.19     0.85
## extraversion        0.5377     2.5047    0.21     0.83
## conscientiousness   1.0266     3.0212    0.34     0.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 115.805  on 105  degrees of freedom
## Residual deviance:  84.538  on 100  degrees of freedom
## AIC: 96.54
##
## Number of Fisher Scoring iterations: 4
```

```
car::Anova(m)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: m_timer
##                   LR Chisq Df Pr(>Chisq)
## log(word_count)       0.16  1       0.69
## project_age          24.71  1    6.7e-07 ***
## openness              0.04  1       0.85
## extraversion          0.05  1       0.83
## conscientiousness     0.12  1       0.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
DescTools::PseudoR2(m)
```

```
## McFadden
##     0.27
```

The VIF index confirm that there are no collinarearity issues as all the scores are < 4.

```
car::vif(m)
```

```
##   log(word_count)       project_age          openness      extraversion
##              1.18              1.12              2.07              2.25
## conscientiousness
##              1.66
```

Finally, we compute the AUC to assess the goodness of the model at predicting whether a one-time contributor will become a contributor (i.e., will keep on contributing afterwards).

```
trainIndex <-
  caret::createDataPartition(p.aggr$m_timer,
                             p = .7,
                             list = FALSE,
                             times = 1)
p.aggr.train <- p.aggr[trainIndex,]
p.aggr.test  <- p.aggr[-trainIndex,]

drops <- c("uid", "one_timer")
p.aggr.train = p.aggr.train[, !(names(p.aggr.train) %in% drops)]
p.aggr.test = p.aggr.test[, !(names(p.aggr.test) %in% drops)]

m1 = stats::glm(
  m_timer ~ log(word_count) +
    openness +
    #agreeableness +
    #neuroticism +
    extraversion +
    conscientiousness,
  data = p.aggr.train,
  family = binomial(link = 'logit')
)
mp <-
  stats::predict.glm(m1,
                     newdata = p.aggr.test,
                     type = "response",
                     se.fit = TRUE)
#mean(mp$fit)
mp1 <- predict(m, newdata = p.aggr.test, type = "response")
mpr <- prediction(mp1, p.aggr.test$m_timer)
mprf <- performance(mpr, measure = "tpr", x.measure = "fpr")
#plot(mprf)

auc <- performance(mpr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

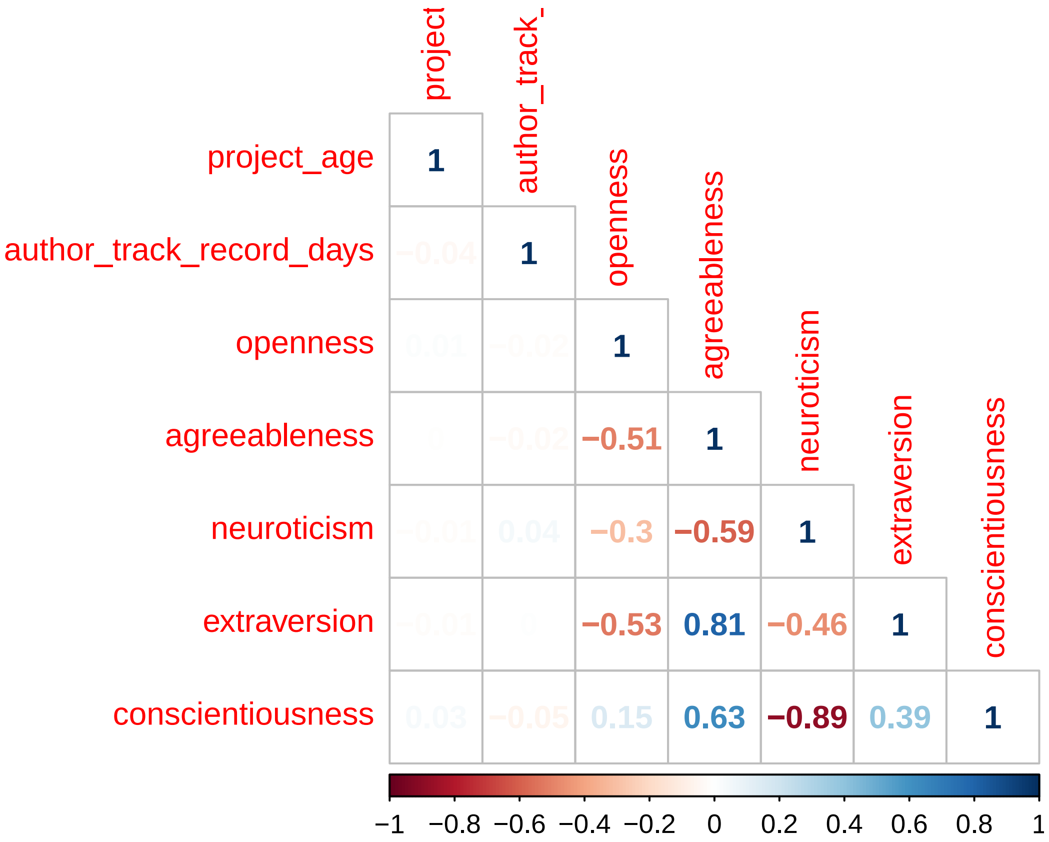
```
## [1] 0.845
```

RQ6 panel data regression model of contribution intensity

In this count data model, the predicted variable is the number of commit per developer.

We prepare the data accordingly.

We repeat again the correlation analysis. We find again the same high correlations between *extraversion* and *agreebleness* , and *conscientiousness* and *neuroticism*. We keep the former from each couple because they give better VIF scores as shown below.



For a correct count data model analysis, we compare two regression strategies, i.e., Poisson and Negative Binomial and choose the better one according to the Likelihood Ratio Test.

```
mod_poisson <- glm(
  commits_merged ~
      log(word_count)
    + is_integrator
    + (scale(project_age, center=TRUE, scale = TRUE))
  + (scale(author_track_record_days, center=TRUE, scale = TRUE))
  + (scale(openness, center=TRUE, scale = TRUE))
  + (scale(conscientiousness, center=TRUE, scale = TRUE))
  + (scale(extraversion, center=TRUE, scale = TRUE))
  #+ (scale(agreeableness, center=TRUE, scale = TRUE))
  #+ (scale(neuroticism, center=TRUE, scale = TRUE))
  , data = panel_data,
  family = "poisson"
)

mod_negbin <- glm.nb(
  commits_merged ~
    log(word_count)
  + is_integrator
  + (scale(project_age, center=TRUE, scale = TRUE))
  + (scale(author_track_record_days, center=TRUE, scale = TRUE))
  + (scale(openness, center=TRUE, scale = TRUE))
  + (scale(conscientiousness, center=TRUE, scale = TRUE))
  + (scale(extraversion, center=TRUE, scale = TRUE))
  #+ (scale(agreeableness, center=TRUE, scale = TRUE))
  #+ (scale(neuroticism, center=TRUE, scale = TRUE))
  , data = panel_data
)
lmtest::lrtest(mod_poisson, mod_negbin)
```

```
## Likelihood ratio test
##
## Model 1: commits_merged ~ log(word_count) + is_integrator + (scale(project_age,
##     center = TRUE, scale = TRUE)) + (scale(author_track_record_days,
##     center = TRUE, scale = TRUE)) + (scale(openness, center = TRUE,
##     scale = TRUE)) + (scale(conscientiousness, center = TRUE,
##     scale = TRUE)) + (scale(extraversion, center = TRUE, scale = TRUE))
## Model 2: commits_merged ~ log(word_count) + is_integrator + (scale(project_age,
##     center = TRUE, scale = TRUE)) + (scale(author_track_record_days,
##     center = TRUE, scale = TRUE)) + (scale(openness, center = TRUE,
##     scale = TRUE)) + (scale(conscientiousness, center = TRUE,
##     scale = TRUE)) + (scale(extraversion, center = TRUE, scale = TRUE))
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   8  -1340
## 2   9  -1023  1   635     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find that Model 2 (Negative Binomial) fits better and has not multi-collinearity issues:

```
car::vif(mod_negbin)
```

```
##                                              log(word_count)
##                                                         1.06
##                                                is_integrator
##                                                         1.06
##              scale(project_age, center = TRUE, scale = TRUE)
##                                                         1.01
## scale(author_track_record_days, center = TRUE, scale = TRUE)
##                                                         1.06
##                 scale(openness, center = TRUE, scale = TRUE)
##                                                         1.77
##        scale(conscientiousness, center = TRUE, scale = TRUE)
##                                                         1.50
##             scale(extraversion, center = TRUE, scale = TRUE)
##                                                         2.04
```

Here is a report of the Negative Binomial model fit. We observe that the only predictors related to personality with a significant effect is *conscientiousness* (coefficient=0.123, p<0.05). Hence, the more organized developers make more commits. Regarding the control variables, we observe that the authors' track record (i.e., the number of days between their first and last successful contribution) has a positive and significant association (coefficient=0.566) with the number of their merged contributions (p<0.001). Similarly, we find a positive and significant association between the response variable and the fact that a developer is a core team member who has integrated external contributions (coefficient=0.474, p<0.05). Instead, project age has a significant, negative effect (-0.087, p<0.05), so the longer the project history, the harder it is to make more commits. However, the model fits the data marginally (Pseudo-R2=0.109) but this was expected as we did not aim for model completeness but rather at understanding the effects of personality traits.

```
summary(mod_negbin)
```

```
##
```

```
## Call:
## glm.nb(formula = commits_merged ~ log(word_count) + is_integrator + 
##     (scale(project_age, center = TRUE, scale = TRUE)) + (scale(author_track_record_days, 
##     center = TRUE, scale = TRUE)) + (scale(openness, center = TRUE, 
##     scale = TRUE)) + (scale(conscientiousness, center = TRUE, 
##     scale = TRUE)) + (scale(extraversion, center = TRUE, scale = TRUE)), 
##     data = panel_data, init.theta = 2.282809723, link = log)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -3.489  -0.602  -0.459   0.037   6.388  
## 
## Coefficients:
##                                                              Estimate
## (Intercept)                                                   1.16905
## log(word_count)                                              -0.04168
## is_integratorTRUE                                             0.47421
## scale(project_age, center = TRUE, scale = TRUE)              -0.08658
## scale(author_track_record_days, center = TRUE, scale = TRUE)  0.56618
## scale(openness, center = TRUE, scale = TRUE)                  0.00381
## scale(conscientiousness, center = TRUE, scale = TRUE)         0.12293
## scale(extraversion, center = TRUE, scale = TRUE)             -0.05362
##                                                              Std. Error z value
## (Intercept)                                                     0.17535    6.67
## log(word_count)                                                 0.02370   -1.76
## is_integratorTRUE                                               0.20370    2.33
## scale(project_age, center = TRUE, scale = TRUE)                 0.04196   -2.06
## scale(author_track_record_days, center = TRUE, scale = TRUE)    0.03308   17.11
## scale(openness, center = TRUE, scale = TRUE)                    0.05431    0.07
## scale(conscientiousness, center = TRUE, scale = TRUE)           0.05071    2.42
## scale(extraversion, center = TRUE, scale = TRUE)                0.05830   -0.92
##                                                              Pr(>|z|)    
## (Intercept)                                                   2.6e-11 ***
## log(word_count)                                                 0.079 .  
## is_integratorTRUE                                               0.020 *  
## scale(project_age, center = TRUE, scale = TRUE)                 0.039 *  
## scale(author_track_record_days, center = TRUE, scale = TRUE)  < 2e-16 ***
## scale(openness, center = TRUE, scale = TRUE)                    0.944    
## scale(conscientiousness, center = TRUE, scale = TRUE)           0.015 *  
## scale(extraversion, center = TRUE, scale = TRUE)                0.358    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(2.28) family taken to be 1)
## 
##     Null deviance: 689.61  on 514  degrees of freedom
## Residual deviance: 389.80  on 507  degrees of freedom
## AIC: 2064
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  2.283 
##           Std. Err.:  0.211 
```

```
##
##  2 x log-likelihood:  -2045.967
```

```
car::Anova(mod_negbin)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: commits_merged
##                                                              LR Chisq Df
## log(word_count)                                                   3.3  1
## is_integrator                                                     5.6  1
## scale(project_age, center = TRUE, scale = TRUE)                   4.3  1
## scale(author_track_record_days, center = TRUE, scale = TRUE)    223.5  1
## scale(openness, center = TRUE, scale = TRUE)                      0.0  1
## scale(conscientiousness, center = TRUE, scale = TRUE)             5.5  1
## scale(extraversion, center = TRUE, scale = TRUE)                  0.8  1
##                                                              Pr(>Chisq)
## log(word_count)                                                   0.068 .
## is_integrator                                                     0.018 *
## scale(project_age, center = TRUE, scale = TRUE)                   0.039 *
## scale(author_track_record_days, center = TRUE, scale = TRUE)     <2e-16 ***
## scale(openness, center = TRUE, scale = TRUE)                      0.944
## scale(conscientiousness, center = TRUE, scale = TRUE)             0.019 *
## scale(extraversion, center = TRUE, scale = TRUE)                  0.359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(mod_negbin)
```

```
## [1] 2064
```

```
BIC(mod_negbin)
```

```
## [1] 2102
```

```
DescTools::PseudoR2(mod_negbin, which = "all")
```

```
##        McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
##           0.109           0.102           0.384           0.388           0.326
## VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
##           0.399          -5.019              NA              NA        2063.967
##             BIC          logLik         logLik0              G2
##        2102.165       -1022.984       -1147.620         249.274
```