Bao Nhi Tang 26004
Thao Nghi Le Thai 26041

BE603: Data Analytics III

Unlock the secrets to crowdfunding success

This presentation aims to address the question often faced by modern entrepreneurs, campaigners and investors:

“What antecedents are associated to a successful crowdfunding campaign?”

Theoretical background & hypotheses

1. Signalling theory & success factors for crowdfunding projects

The notion of the signalling theory established by Spence (1973, 2002) can be applied to the crowdfunding project. Considering that investors do not know the credibility of the project before funding, the number of posted updates and comments can serve as positive proxies to signal **engagement** & thus trustworthiness (Beier & Wagner, 2015; Kuppuswamy & Bayus, 2017). Within the context of the dataset provided, based on the correlation matrix presented on the next slide, **comments_count** seems to be the most prominent variable of all. Besides, we control for variables indicating project quality. We test this first hypothesis:

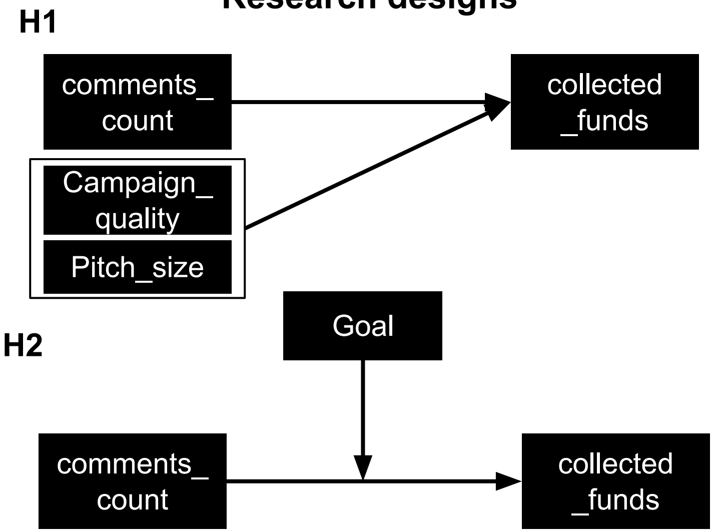
*H1.There is a **positive and significant relationship** between **the collected_funds of crowd-funding projects and the number of comments,** such that projects with more comments_count raise more collected_funds.*

2. Moderating effect of goal on success factors

Frydrych et al. (2014) argue that a high funding goal can decrease the legitimacy of the project if the project founders do not provide a convincing justification for the set funding goal. If that is true, proof of positive success factors, such as engagement, becomes evermore important to success as goal increases; while projects with lower goal does not need as much evidence to achieve what they want. We test our second hypothesis using comments_count as a proxy for success factor engagement:

*H2. **The goal level has a moderating effect on the relationship** between the **collected_funds** of crowdfunding projects and the **comments_count of the project** such that the effect of comments_count on collected fund is larger for projects with higher goal.*

Research designs



Data & methodology

This study analyzes a dataset of 5000 crowdfunding campaigns from a US-based platform. Here are a list of variables being used in the report:

- ***Dependent variable: collected_funds***

The total of collected_funds by the end of the crowdfunding campaign, in dollars

- ***Independent variable: comments_count***

The number of comments posted by the backers on the campaign page by the end of the campaign

- ***Moderator: goal***

The total amount that the entrepreneurs aimed to raise during the campaign

- ***Control variables: pitch_size & campaign_quality***

Pitch_size: the number of words that appear in the pitch written on the campaign page

Campaign_quality: Campaign quality is a composite variable that has a higher value when the campaign page includes tokens of quality such as images, videos, or perks.

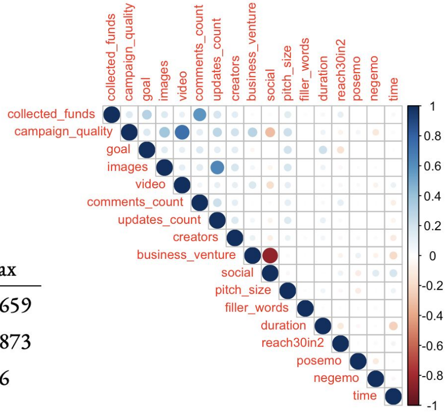


Fig 1. Correlation matrix & descriptive statistics

Summary Statistics

Variable	Nr. Observation	Mean	S.D.	Min	Max
collected_funds	5000	2862	3572	499	77659
goal	5000	6185	4997	1000	24873
comments_count	5000	16.8	24.6	0	886

The first hypothesis is tested using a linear OLS-regression of collected funds on comments_count and control variables pitch_size & campaign_quality, making sure we compare projects with similar proxies of quality.

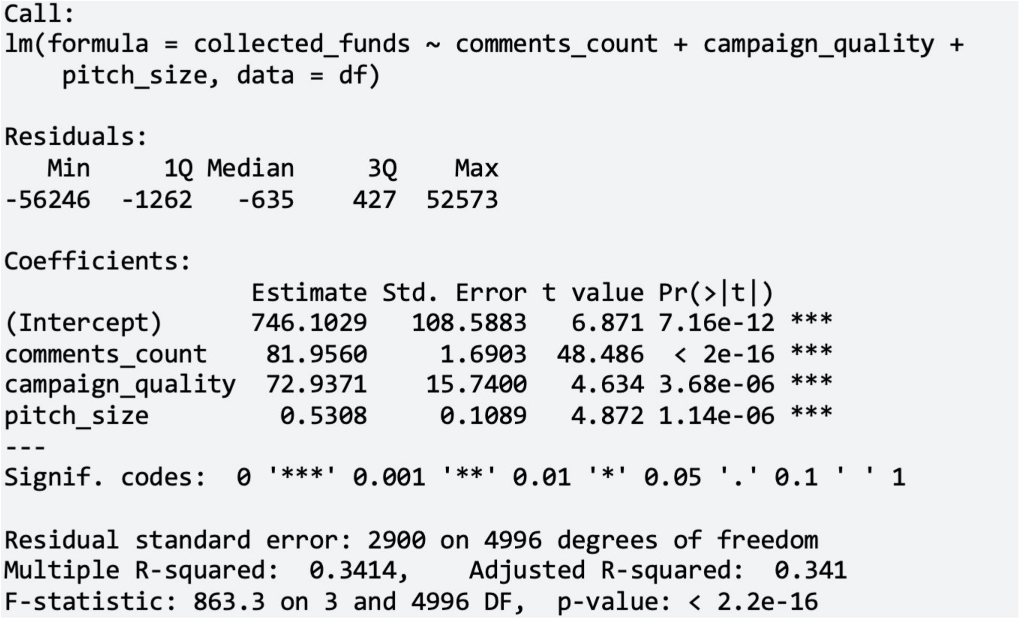
The second hypothesis tested the potential moderating effect of different funding goal levels for comments_count on collected_funds. Two approaches were used:

- **Goal category approach:** We employ linear OLS-regression and separate four levels of funding goals. The assessed projects were assigned to four categories determined by three different funding goal thresholds: The 25%-percentile of the funding goal in our dataset at $2500, the 50%-percentile at around $5000, and the 75%-percentile at $9500.

- **All data approach:** We employ linear regression of collected_funds on comments_count with goal as a multiplicative term in the linear equation, without separating projects with respect to goal levels. The model is then tested for outliers and refit using the cleaned dataset. No standardization of data since preliminary testing reveals little multicollinearity.

Testing H1

The coefficient of comments_count on collected_funds is 81.96, which is positive and significant, given the control variables being pitch_size and campaign_quality. This implies that with the same level of campaign_quality and same pitch_size, one unit increase in comments_count would result in 81.96 additional dollars in the collected_funds. The R-squared implies that 34.14% of the collected_funds can be explained by the number of comments. The p-values for all coefficients & for the models are below 0.001. Therefore, we reject the null hypothesis, with the alternative hypothesis saying that there is a positive and significant relationship between the collected_funds of crowdfunding projects and the number of comments_count, such that projects with more comments_count can raise more funds.



```
Call:
lm(formula = collected_funds ~ comments_count + campaign_quality +
    pitch_size, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-56246  -1262   -635    427  52573

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      746.1029   108.5883   6.871 7.16e-12 ***
comments_count    81.9560     1.6903  48.486  < 2e-16 ***
campaign_quality  72.9371    15.7400   4.634 3.68e-06 ***
pitch_size         0.5308     0.1089   4.872 1.14e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2900 on 4996 degrees of freedom
Multiple R-squared:  0.3414,	Adjusted R-squared:  0.341
F-statistic: 863.3 on 3 and 4996 DF,  p-value: < 2.2e-16
```

Output 1. OLS-regression of collected_funds on comments_count

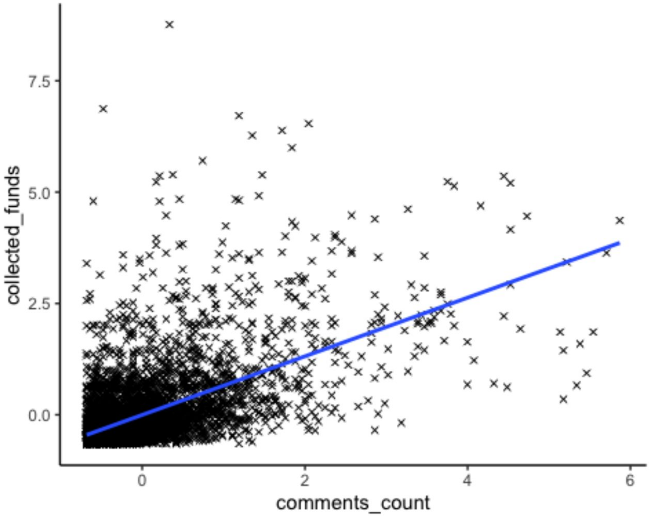
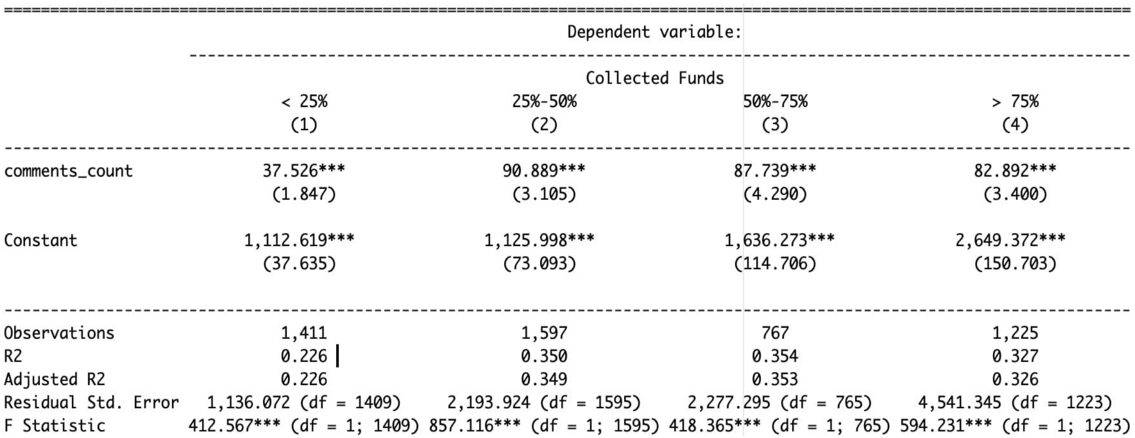


Fig 2. Relationship between collected_funds and comments_count

Testing H2 - Goal quartile approach



```
==========================================================================================================
                                          Dependent variable:
                 -----------------------------------------------------------------------------------------
                                            Collected Funds
                          < 25%                  25%-50%                 50%-75%                > 75%
                           (1)                     (2)                     (3)                    (4)
----------------------------------------------------------------------------------------------------------
comments_count          37.526***              90.889***               87.739***              82.892***
                         (1.847)                (3.105)                 (4.290)                (3.400)

Constant              1,112.619***           1,125.998***            1,636.273***           2,649.372***
                        (37.635)               (73.093)               (114.706)              (150.703)

----------------------------------------------------------------------------------------------------------
Observations              1,411                  1,597                    767                   1,225
R2                        0.226                  0.350                   0.354                  0.327
Adjusted R2               0.226                  0.349                   0.353                  0.326
Residual Std. Error  1,136.072 (df = 1409)   2,193.924 (df = 1595)   2,277.295 (df = 765)   4,541.345 (df = 1223)
F Statistic         412.567*** (df = 1; 1409) 857.116*** (df = 1; 1595) 418.365*** (df = 1; 765) 594.231*** (df = 1; 1223)
```

Output 2. OLS-regressions of collected_funds on comments_count with at 4 quartiles of goal

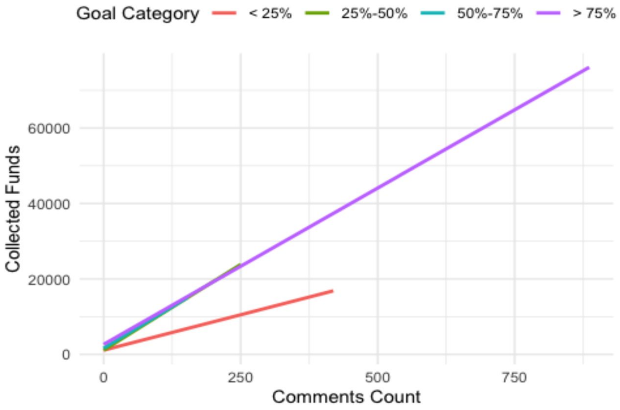
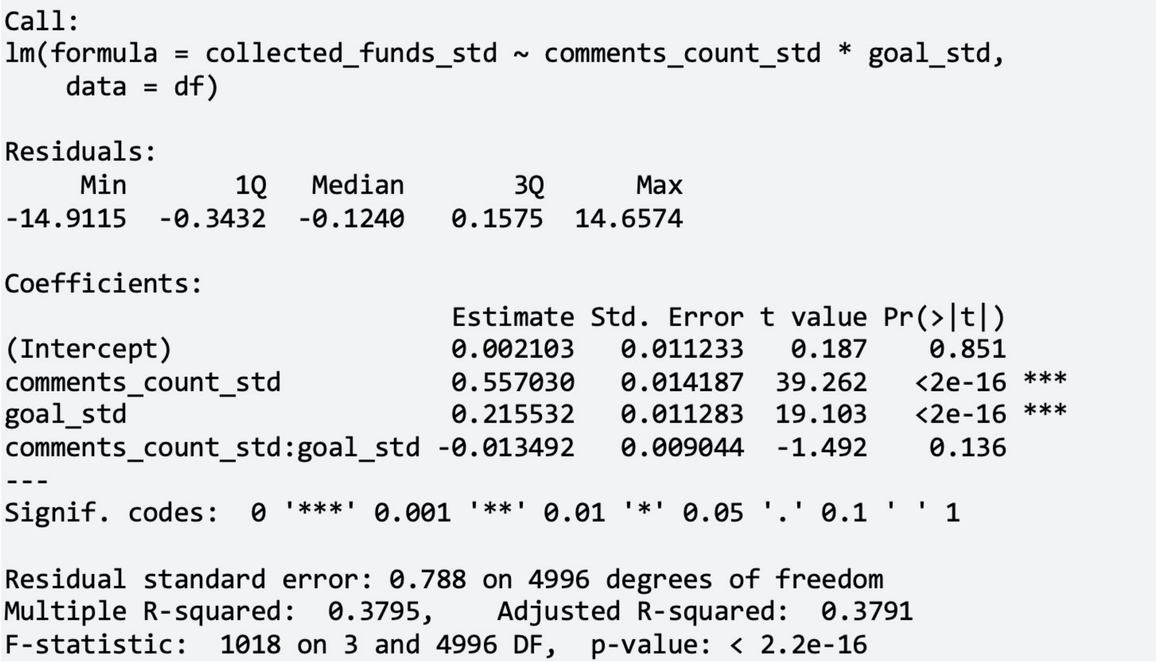


Fig 3. The correlation of comments_count versus collected_funds at different goal level

The compiled results show a significant p-value for every goal quartile, further supporting the hypothesis that comments_count is positively correlated with collected_funds. The **moderating effect of goal** is also clear through the different interactions of the model across different quartile. It can inferred from the output that at the 25% quartile (goal < $2500), one unit increase in comments_count only implies about $37.53 added to the fund, while this number is > $80 for all larger quantiles. The R-squared of the OLS-regression is also significantly smaller for projects will goal <$2500, with comments_count explaining only 22.6% of variation in data, whereas this measure peaks at 35% for projects whose goal ranges from $5000 to $9500 (3rd quartile). *The reasoning behind this result is that projects with higher goal would need engagement (comments) to signal credibility & therefore comments influence success more.*

However, when the goal level is above 75%-percentile, both the coefficient and the explanatory power start to slightly lower down. There could be a possibility that when the goal is set too high, the amount of collected_funds will be slightly less dependent on the comments_count, meaning that other control variables can further explain the relationship.

Testing H2 - All data approach



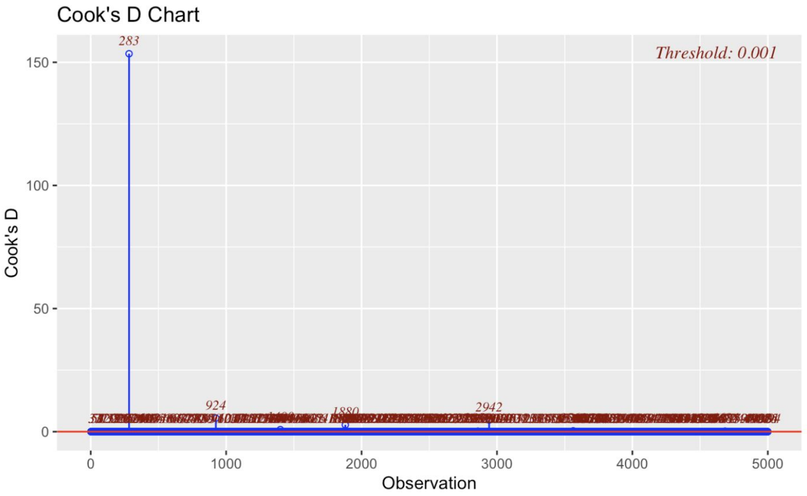
```
Call:
lm(formula = collected_funds_std ~ comments_count_std * goal_std,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9115  -0.3432  -0.1240   0.1575  14.6574

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.002103   0.011233   0.187    0.851
comments_count_std           0.557030   0.014187  39.262   <2e-16 ***
goal_std                     0.215532   0.011283  19.103   <2e-16 ***
comments_count_std:goal_std -0.013492   0.009044  -1.492    0.136
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.788 on 4996 degrees of freedom
Multiple R-squared:  0.3795,	Adjusted R-squared:  0.3791
F-statistic:  1018 on 3 and 4996 DF,  p-value: < 2.2e-16
```

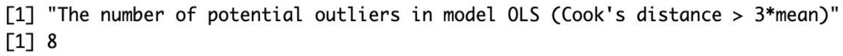
Output 3. Regression of collected _funds on comments_count with goal being the moderator

The coefficients for comments_count and goal are 0.557030 and 0.215532 respectively, suggesting that one standard deviation increase in comments_count results in an additional of 0.557030 standard deviation increase in collected_funds and one standard deviation increase in goal results in an additional of 0.215532 standard deviation increase in collected_funds. The p-value for both coefficients are significant while the p-value of the interaction term is 0.136, inferring that the interaction between goal and comments_count is **not statistically significant** in predicting the collected_funds. The model can explain 37.95% (R-squared) of the variance in the dependent variable, which is collected_funds.

All variables are standardized

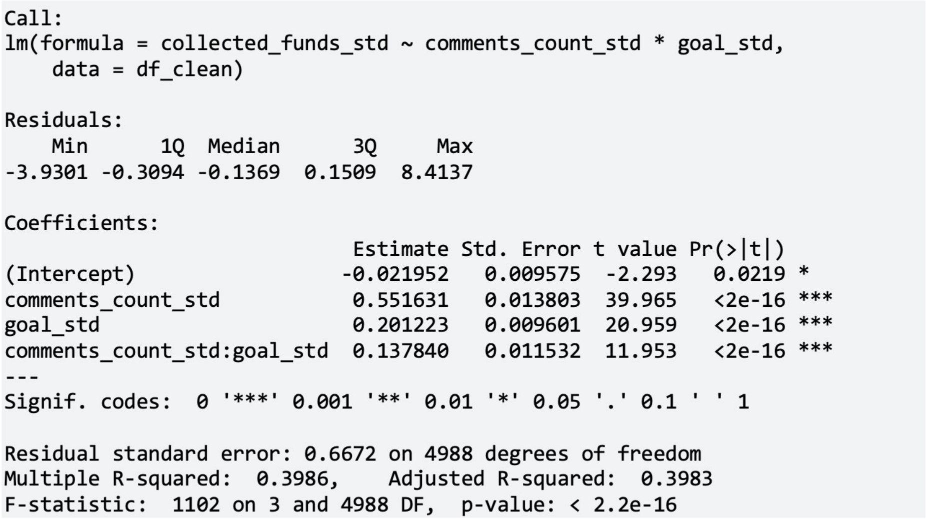
Testing H2 - All data approach_refit





```
[1] "The number of potential outliers in model OLS (Cook's distance > 3*mean)"
[1] 8
```

Output 4. Cook’s D Chart indicating potential outliers to the model



```
Call:
lm(formula = collected_funds_std ~ comments_count_std * goal_std,
    data = df_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9301 -0.3094 -0.1369  0.1509  8.4137 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -0.021952   0.009575  -2.293   0.0219 *  
comments_count_std           0.551631   0.013803  39.965   <2e-16 ***
goal_std                     0.201223   0.009601  20.959   <2e-16 ***
comments_count_std:goal_std  0.137840   0.011532  11.953   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6672 on 4988 degrees of freedom
Multiple R-squared:  0.3986,	Adjusted R-squared:  0.3983 
F-statistic:  1102 on 3 and 4988 DF,  p-value: < 2.2e-16
```

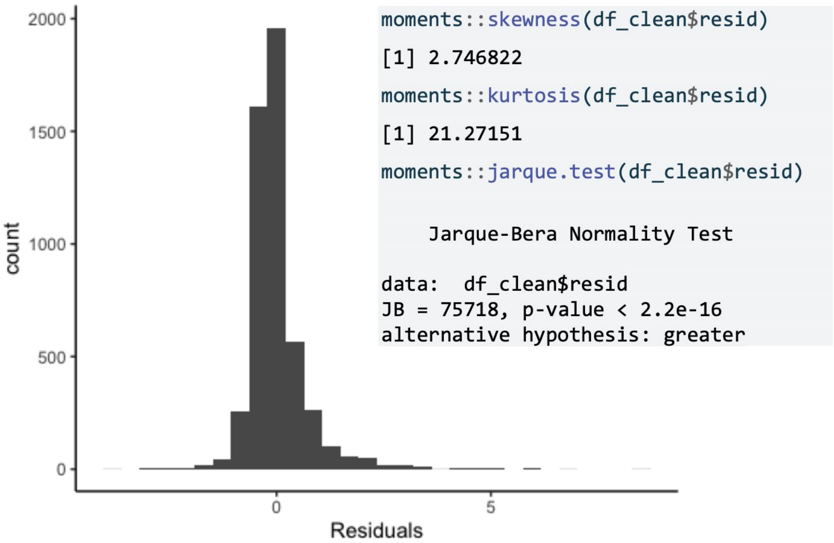
Output 5. OLS-regression of collected_funds on comments_count with goal as the moderator, outliers removed

After testing for outliers using the Cook’s distance, 8 potential outliers were identified and removed from the data. A refitted model was run again.

The coefficients for comments_count and goal, as well as the interaction term is significant (p-value<0.001), supporting the hypothesis that goal has a moderating effect on the relationship between comments_count and collected_funds. The sign of the interaction term is positive, indicating that higher goals amplify the effect of comments_count on the collected_funds.
The model can explains 39.86% of variation in data (R-squared = 0.3986).

All variables are standardized

Checking Assumptions



```
moments::skewness(df_clean$resid)
[1] 2.746822
moments::kurtosis(df_clean$resid)
[1] 21.27151
moments::jarque.test(df_clean$resid)


	Jarque-Bera Normality Test

data:  df_clean$resid
JB = 75718, p-value < 2.2e-16
alternative hypothesis: greater
```

Fig 4. Normality of error terms check

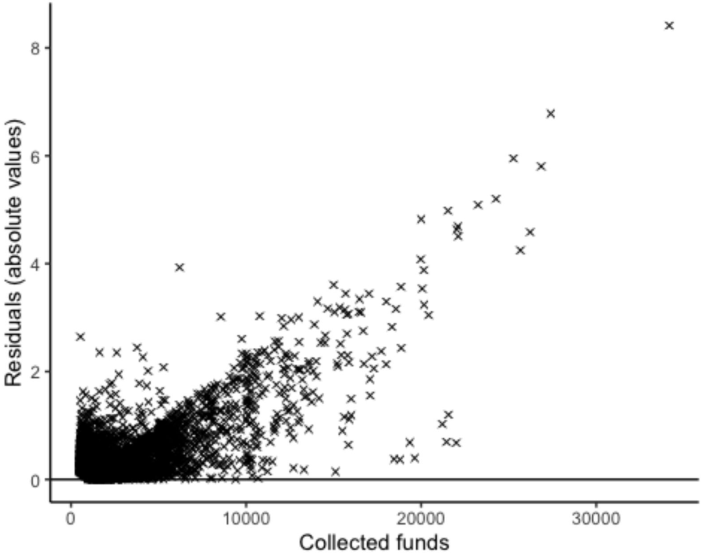
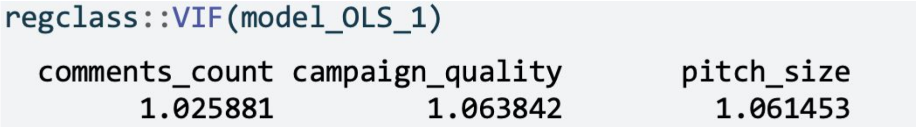


Fig 5. Normality of error terms check

The distribution of error terms is not normal. The skewness and kurtosis are 2.746822 and 21.27151, suggesting the the distribution is positively skewed to the right and has heavy tails. This means that the model is slightly biased. Thus, the standard error of OLS estimates is not reliable, causing the confidence intervals to be too narrow. The model may give over-optimistic impression of the precision; in fact, uncertainty does exist.

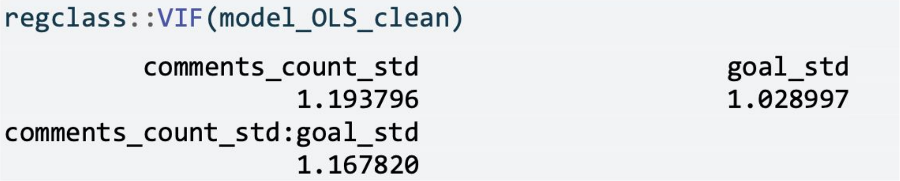
Heteroskedasticity noticed for error terms, indicating that there are confounding factors behind collected_funds besides from comments_count. The variance of error terms are relatively stable at collected fund below $10,000, indicating that the model might be more efficient at predicting lower fund levels. For a larger amount of fund to be collected, complex confounding factors must come into play to explain the success.

Checking Assumptions



```
regclass::VIF(model_OLS_1)

  comments_count campaign_quality        pitch_size
        1.025881         1.063842          1.061453
```

Output 6. Multicollinearity check - hypothesis 1



```
regclass::VIF(model_OLS_clean)

           comments_count_std                      goal_std
                     1.193796                      1.028997
comments_count_std:goal_std
                     1.167820
```

Output 7. Multicollinearity check - hypothesis 2

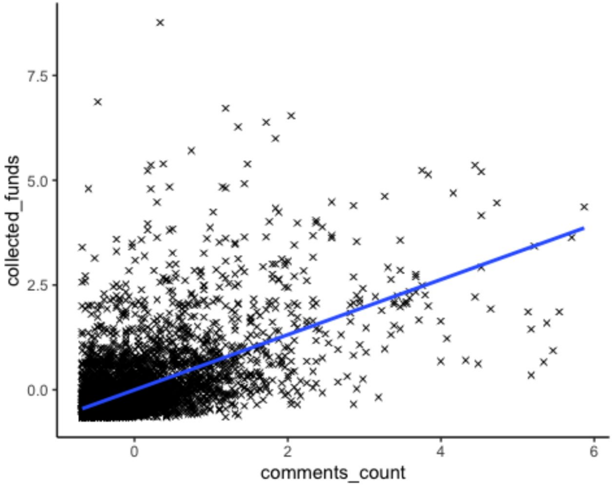


Fig 6. Linearity check

VIF values for comments_count, campaign_quality and pitch_size are approximately 1.06, 1.03 and 1.6 respectively, showing that there is very low or no correlation among the variables.

Similarly, VIF values for comments_count, goal and the interaction term between them are approximately 1.19, 1.03 and 1.17 respectively, showing that there is very low or no correlation among the variables.

The relationship between collected_funds and comments_count is approximately linear. Meanwhile, there is some variability in the data points around the line, particularly at higher values of comments_count. In general, the relationship is till considered as linear, as the blue line of best fit is well-fitted in the data, showing an upward trend.

Conclusion

After testing the two hypotheses, it can be concluded that:

H1 is supported, showing that *there is a positive and significant relationship between* ***the collected_funds of crowd-funding projects and the number of comments,*** *such that projects with more comments_count raise more collected_funds.*

H2 is supported, showing that ***the goal level has a moderating effect on the relationship*** *between the* ***collected_funds*** *of crowdfunding projects and the* ***comments_count of the project*** *such that the effect of comments_count on collected fund is larger for projects with higher goal.*

These results demonstrate the importance of engagement (particularly via comments) in the success of crowdfunding. Regarding the moderating effect, the increasing levels of goal will result in stronger effects on the relationship between comments_count and collected_funds. However, when the goal is set too high (above 75%-percentile), collected_funds is less likely to depend on comments_count as the larger projects mean less backers' individual contribution in proportion to the funding goal, reducing the perceived impact of individual interactions such as comments_count. Hence, our hypothesis highlights that entrepreneurs and investors should have adaptive strategies depending on the size of funding goal.

However, when checking the assumptions of the models, two assumptions about normality of error terms and homoscedasticity are not met. This would require further statistical techniques to address these issues.

By proposing and testing the two hypotheses, the interplay between social engagement by comments and strategic goal-setting in crowdfunding projects is clearly justified and supported, thus, posing effects on the success of the projects. The findings highlight one of the key drivers for crowdfunding success and practical guidance for entrepreneurs and investors on tactical objectives setting.