Vladimir Averin, gr.1., vsaverin@edu.hse.ru, Applied essay, year 3,
April 2022, kickstarter data, python (Jupyter notebook).
Estimation for the optimal number of words in the description of the Gaming Project on Kickstarter

1 Introduction

(I note that all the econometric work was done in python (Jupyter Notebook) in the file:

Averin_applied_essay_econometrics_gr1_code_draft.ipynb, attached to the zip file, so all my work can be observed there and some screenshots from this file will be included in the appendix.)

In this essay I will consider gaming projects published on the site Kickstarter.com. This is the site where everyone may publish their own project with description, set a goal which is the sum needed to be raised for the project to be realised and then all the users of the Kickstarter.com may invest in the project.

I am going to analyse, how the number of words in description description affects the amount of money the gaming project raised. My hypothesis is as follows: if description of the project is not long enough then this project may seem to be quite raw and in the eyes of investors and the game developers may not be trusted, so the project is going to raise not a lot of money. On the other hand, if the description is too long, investors may find irrational to investigate the project (time-consuming to read long description), so they would rather find projects with shorter descriptions, so I think that in this case there should exist an optimal number of words in the description of the project, that will attract the highest number of investors.

I suppose that this analysis and estimation of the optimal number of words will help to make fund-raising of the gaming project on Kickstarted.com more efficient. For example, suppose if investors had been received two finished games: A and B. Investor played these games and decided that investor liked game A more than game B. Now let's return to reality, where investor sees only the Kickstarter pages of both games. Investor observes that description of the game A's description on Kickstarter is too short or too long, which confuses investor and because of that investor invests in game B, but he would have been preferred game A over game B. So if optimal number of words is known, then all game developers will adjust their games' description to the 'standard' word count, so the situations like the one above with investor and games A and B will be less likely to happen. So my analysis will help game developers to adjust their projects, so that investors will be more likely to invest in the projects which they will like the most, so investors will be more satisfied. And increased satisfaction of investors in turn may stimulate investors to invest in gaming projects more funds, so game developers and Kickstarter itself will benefit from that as well.

I note that all the econometric work was done in python (Jupyter Notebook) in the file:

Averin_applied_essay_econometrics_gr1_code_draft.ipynb, attached to the zip file, so all my work can be observed there and some screenshots from this file will be included in the appendix.

2 Plan

So plan of the essay is as follows:

3) Methodology: I will briefly outline how I am going to estimate the optimal number of words in the Kickstarter description

4) Data: I will describe, where I got the data and which variables it has initially and where this variable is included in the analysis

5) Preparation of the variables to be used in the analysis: Transformation of the initial variables which may lead to the improvement of the model estimation

6) Model selection: out of all possible candidates for the optimal model, I will choose "the one and only"

7) Testing the model for "validity": since the data is cross-sectional, I am going to test my chosen model for consistency of parameters' estimates and heteroscedasticity only.

8) Interpretation of the model and estimation of the optimal number of words

9) Sample distribution of the optimal number of words using Bootstrap method, calculating confidence intervals for this parameter

10) Conclusion

11) Further analysis: Some things which can extend the analysis and/or make it of the better quality

12) Bibliography

13) Appendix: some technical information which may not demonstrated in the essay, but should be provided to support the statements of the essay.

3 Methodology

To find the optimal number of the words in the description I first need to get the valid representation of the amount of funds raised as a function of the number of the words in the project: $funds\_raised = f(number\_of\_text\_in\_the\_description)$. So, we may just use the number of words and some other control variables and regress amount of funds raised on this set of variables, right?

Not quite, we should add the number of words in the description squared, since otherwise the function $f(number\_of\_text\_in\_the\_description)$ is linear and thus, the optimal number of words will be either 0 or infinitely big number (depending on the sign of the parameter to the number of words variable). Adding squared variable helps to make the function $f(number\_of\_text\_in\_the\_description)$ parabolic and thus if the coefficient to the squared number of words in the description is negative, then function $f(number\_of\_text\_in\_the\_description)$ has the point of maximum and thus this point will be exactly the optimal number of words in the description of the project. (Of course, if such optimal number is within some reasonable range, i.e. min and max number of words of description of the whole data) (of course, we also should have the list of control variables, since otherwise, there will be omitted variable bias and the estimates of coefficients will be inconsistent.

Note that if coefficient to the number of the words in description is not statistically significant or even significantly positive then optimal number of words does not exits and thus my initial hypothesis about existence of optimal number of words in the description is false, so the whole analysis becomes invalid.

4 Data

4.1 Description of the initial dataset

Initial data which I am going to use contains information parsed directly from Kickstarter.com, which was created directly for the Data Analysis National Olympiad (DANO) for secondary school students which was help in December 2021 (you can find this data in the file kick_data_nice_final.xlsx in the submitted zip file). After the parsing the following variables were added in the model: URL, pled, goal, date, period, Status, text_am, n_vid, n_img, game categories variables (rpg, platformer, shooter, fighting, survival, horror, strategy, arcade, simulator, mmo, indie, action, quest, adventure, mgp), n_pled_t, min_pled_t, step_pled_t, cr_time, backed, created, mgp, cont, curr, succsess and site. The meaning of these variables can be found in the Appendix 13.1! Out of these variables, only succsess, URL and site will not be used of mentioned in my analysis. So the target variable will be derived from pled and the explanatory models will be transformed from the other variables which will be used in my analysis.

4.2 Target variable

Since my topic is to observe the effect of number of words in the description (text_am) on the funds raised by the project, it is logical to use pled as a target variable. However, it is logical that the difference between the project which raised 10000 USD and the one which raised 100 USD is massive, while the difference between the project which raised 1 000 000 and the one which raised 990 100 is not quite large, but if simple pled is used, the regression will treat these differences in the same way. That is why I am going to log the variable pled and use log_pled as the target variable.

5 Preparation of the variables to be used in the analysis

5.1 filtering the data and throwing out some variables, which will not be needed in the further analysis

First of all, initially there were 8421 gaming projects, but not all projects are suitable for my analysis. As 70% of projects in the data raise funds in USD (variable curr), I decided to filter out all other currencies, to avoid the heterogeneity of fund-raising in different currencies. Also I deleted the several projects which did not stop raising funds as at the moment of parsing (variable status) (Note, that variables curr and status will not be used later on).

Also it was observed that mean_pled_t is highly correlated with step_pled_t and max_pled_t which is quite logical. Because of that and the fact that min/max/step_pled_t already may explain the donation options of the project quite well, I decided to exclude mean_pled_t from the further analysis! (Also this deletion may help to reduce multicollinearity in the data)

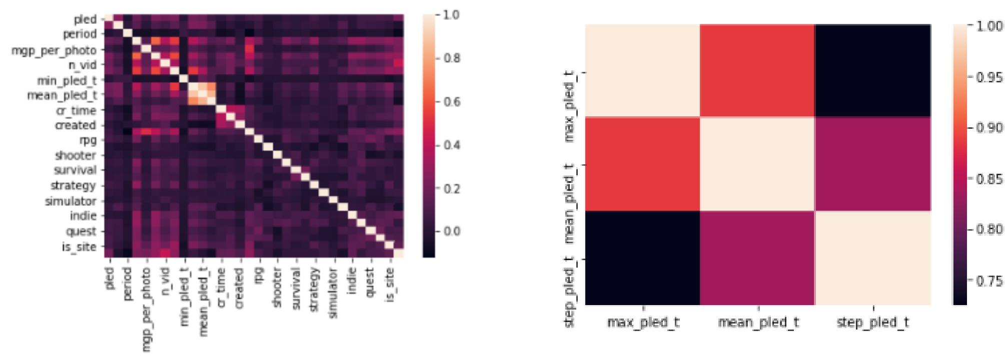

Figure 1: Correlation between variables in the data (colour indicates correlation coefficient between variables). The picture on the right shows the only set of variables with high correlation which may be the source of multicollinearity, that is one of the reasons why variable mean_pled_t was deleted

Finally, I am going to filter out data before 2012, because it at the period before 2012 Kickstarter was not as popular as after 2012 and more importantly, it was not considered seriously as a fund-raising platform, so data from this period may reduce the quality of the analysis

These are the only filters of the projects, so the filtered data contains 5379 projects.

5.2 Variable transfomation

This section will be devoted to the creation of new variables from the initial explanatory variables which may act as an additional or alternative explanatory variables to some set of existing explanatory variables.

First of all, I am going to create text_am_sq variable: the number of words in the description SQUARED, since, as I have mentioned earlier, this is the key variable of my analysis.

Secondly, we have mgp which is the sum of all pixels of all photos in the description, but it is obvious that it is correlated with the number of photos (n_img variable), so I suppose that creation mgp_per_photo variable makes more sense, because this variable will show the average resolution of the game, so this the creation of another factor of amount of funds raised: higher-resolution photos may make the description of the project more attractive to the investor, so the game project is likely to raise more funds. Thus, I substitute mgp onto mgp_per_photo.

Furthermore, after some exploratory data analysis (check appendix 13.2 for the details), the following alternative variables were created (there are 4 sets in the list below, so in appendix I will explain why I decided to create another variables, and I will references to them by the number of set for convenience):

1) log_goal, log_mgp_per_photo, log_n_img, log_cr_time, log_min_pled_t, log_step_pled_t - logarithms of the corresponding variables from the initial data.

2) n_img_binary and n_vid_binary - binary variables which indicate whether the description has photo/video (1 if has and 0 if doesn't have)

3a) tags_sum - number of tags the project has in its description

3b) tags_sum1_1, tags_sum1_2, tags_sum1_3, tags_sum1_¿3 - Binary variables which indicate the number of the tags the game has (1, 2, 3 or more than 3, and no tags is reference category)

4a) date_timestamp - number of seconds passed from year 0, so this variable indicates time variable.

4b) date_year_2013, date_year_2014, date_year_2015, date_year_2016 - dummy variables which indicate the year of the publication of the project on Kickstarter (year 2012 is the reference category)

So I listed 4 SETS (log, not log; binary image/video; tags; date) where I can choose the best alternative and I am going to do that in the next section.

6 Model selection

6.1 Algorithm

So as I've mentioned in the last section, I have 4 sets, where I can choose the best alternative group and, thus the optimal model will be created.

More specifically, the approach is as follows:

1) Start with default model:

$log\_pled_i = \beta_0 + \beta_1 text\_am_i + \beta_2 n\_img_i + \beta_3 created_i + \beta_4 mgp\_per\_photo_i + \beta_5 quest_i + \beta_6 max\_pled\_t_i + \beta_7 is\_site_i + \beta_8 mmo_i + \beta_9 backed_i + \beta_{10} strategy_i + \beta_{11} simulator_i + \beta_{12} period_i + \beta_{13} adventure_i + \beta_{14} n\_pled\_t_i + \beta_{15} text\_am\_sq_i + \beta_{16} min\_pled\_t_i + \beta_{17} shooter_i + \beta_{18} n\_vid_i + \beta_{19} horror_i + \beta_{20} arcade_i + \beta_{21} survival_i + \beta_{22} platformer_i + \beta_{23} cr\_time_i + \beta_{24} step\_pled\_t_i + \beta_{25} cont_i + \beta_{26} indie_i + \beta_{27} fighting_i + \beta_{28} rpg_i + \beta_{29} goal_i + \beta_{30} action_i + \beta_{31} date\_timestamp + v_i$

2) Considering one of the four sets (last section) and regress all the possible alternatives (using OLS method), so basically I run default regression and then substitute the set of initial variables FROM THE ONE OF THE FOUR SETS onto the alternative variables. For example, the default model contains non-log variables (goal, mgp_per_photo, n_img, cr_time, min_pled_t, step_pled_t) and other controls. I run this regression and then I substitute non-log variables onto log counterparts and run this regression as well. Then I compare these regressions (using information criterions and where possible, I use Zarembka test to find out whether one model is significantly better than the other)

3) As better model this chosen I set this model as default and repeat the process from step 2), until 4 sets are considered.

4) Removing insignificant variables.

Below I will state the optimal variables in each set and in the Appendix 13.3 I will show the detailed explanations why I have chosen certain variables:

a) Set 1: (goal, mgp_per_photo, n_img, cr_time, min_pled_t, step_pled_t) VS (log_goal, log_mgp_per_photo, log_n_img, log_cr_time, log_min_pled_t, log_step_pled_t): it is better to use log variables

b) Set 2a: (n_vid_binary) VS (n_vid): it is better to use binary variable

Set 2b: whether n_img_binary is significant: No, this variable is not significant

c) Set 3: (rpg, platformer, shooter, fighting, survival, horror, strategy, arcade, simulator, mmo, indie, action, quest, adventure, mgp) VS (tags_sum) VS (tags_sum1_1, tags_sum1_2, tags_sum1_3, tags_sum1_¿3): better to use (rpg, platformer, shooter, fighting, survival, horror, strategy, arcade, simulator, mmo, indie, action, quest, adventure, mgp)

d) Set 4: (date_timestamp) VS (date_year_2013, date_year_2014, date_year_2015, date_year_2016): better to use dummy variables.

So the model with the best variables is as follows:

$log\_pled_i = \beta_0 + \beta_1 text\_am_i + \beta_2 log\_n\_img_i + \beta_3 created_i + \beta_4 log\_mgp\_per\_photo_i + \beta_5 quest_i + \beta_6 max\_pled\_t_i + \beta_7 is\_site_i + \beta_8 mmo_i + \beta_9 backed_i + \beta_{10} strategy_i + \beta_{11} simulator_i + \beta_{12} period_i + \beta_{13} adventure_i + \beta_{14} n\_pled\_t_i + \beta_{15} text\_am\_sq_i + \beta_{16} log\_min\_pled\_t_i + \beta_{17} shooter_i + \beta_{18} n\_vid\_binary_i + \beta_{19} horror_i + \beta_{20} arcade_i + \beta_{21} survival_i + \beta_{22} platformer_i + \beta_{23} log\_cr\_time_i + \beta_{24} log\_step\_pled\_t_i + \beta_{25} cont_i + \beta_{26} indie_i + \beta_{27} fighting_i + \beta_{28} rpg_i + \beta_{29} log\_goal_i + \beta_{30} action_i + \beta_{31} date\_year\_2013_i + \beta_{32} date\_year\_2014_i + \beta_{33} date\_year\_2015_i + \beta_{34} date\_year\_2016_i + v_i$

6.2 Deleting/aggregating variables

The process of the deleting/aggregating variables is as follows:

1) Choose set of variable, and made a restriction which allows to determine whether the variable(s) can be deleted/aggregated (whether the restriction is valid)

2) Update the model according to the restriction (if it is valid) and repeat step 1) until only significant groups of variables are left in the model. So the detailed procedure can be seen in the Appendix 13.4, so the preliminary final model is as follows:

$log\_pled_i = \beta_0 + \beta_1 text\_am_i + \beta_2 log\_n\_img_i + \beta_3 log\_mgp\_per\_photo_i + \beta_4 is\_site_i + \beta_5 mmo_i + \beta_6 backed_i + \beta_7 adventure_i + \beta_8 n\_pled\_t_i + \beta_9 text\_am\_sq_i + \beta_{10} log\_min\_pled\_t_i + \beta_{11} shooter_i + \beta_{12} n\_vid\_binary_i + \beta_{13} log\_cr\_time_i + \beta_{14} log\_step\_pled\_t_i + \beta_{15} indie_i + \beta_{16} fighting_i + \beta_{17} rpg_i + \beta_{18} log\_goal_i + \beta_{19} action_i + \beta_{20} date\_year\_2013_i + \beta_{21} date\_year\_2014\_15\_16_i + v_i$

And the estimation of it:

Dep. Variable:	log_pled	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	342.6
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:09	Log-Likelihood:	-11108.
No. Observations:	5379	AIC:	2.226e+04
Df Residuals:	5357	BIC:	2.240e+04
Df Model:	21		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
date_year_2014_15_16	-0.7172	0.072	-9.930	0.000	-0.859	-0.576
log_min_pled_t	0.1164	0.037	3.138	0.002	0.044	0.189
shooter	-0.1867	0.091	-2.062	0.039	-0.364	-0.009
n_vid_binary	1.6705	0.064	25.986	0.000	1.545	1.797
text_am	0.0013	0.000	10.487	0.000	0.001	0.002
log_cr_time	0.0691	0.016	4.365	0.000	0.038	0.100
log_n_img	0.4581	0.036	12.748	0.000	0.388	0.529
log_mgp_per_photo	1.2928	0.253	5.106	0.000	0.797	1.789
log_step_pled_t	0.1045	0.022	4.687	0.000	0.061	0.148
indie	0.2158	0.061	3.560	0.000	0.097	0.335
is_site	0.4603	0.069	6.694	0.000	0.326	0.595
mmo	-0.3730	0.107	-3.480	0.001	-0.583	-0.163
log_goal	0.2633	0.023	11.692	0.000	0.219	0.307
const	-0.5256	0.185	-2.844	0.004	-0.888	-0.163
fighting	-0.2490	0.077	-3.231	0.001	-0.400	-0.098
backed	0.0086	0.003	2.595	0.009	0.002	0.015
date_year_2013	-0.3267	0.080	-4.092	0.000	-0.483	-0.170
adventure	0.2183	0.059	3.722	0.000	0.103	0.333
n_pled_t	0.0916	0.012	7.844	0.000	0.069	0.114
rpg	0.1576	0.064	2.471	0.013	0.033	0.283
text_am_sq	-2.292e-07	2.69e-08	-8.525	0.000	-2.82e-07	-1.77e-07
action	-0.1168	0.061	-1.912	0.056	-0.237	0.003

Figure 2: Estimation of preliminary final model

7 Testing the model for "validity"

7.1 Checking for the consistency of the estimates

I assume that the model with the best coefficients without deletion/aggregation of the variables is consistent, so I will check the coefficients of final preliminary model for the similarity with the consistent model. Hausman test can be used, but I am not exactly sure how to do it manually (Basically I understand that apart from coefficients I need Var(coef i of Model A - coef i of Model B), but I am not sure how to find that). So let's just analyse and compare the coefficients ourselves:

	final_model_estimates	final_model_std_errors	raw_model_estimates	raw_model_std_errors
log_min_pled_t	1.163938e-01	3.709755e-02	1.095243e-01	3.755305e-02
shooter	-1.866979e-01	9.056368e-02	-1.867036e-01	9.173172e-02
n_vid_binary	1.670512e+00	6.428593e-02	1.665222e+00	6.449746e-02
text_am	1.271364e-03	1.212381e-04	1.255430e-03	1.229736e-04
log_cr_time	6.911183e-02	1.583357e-02	7.106781e-02	1.617370e-02
log_n_img	4.580889e-01	3.593405e-02	4.862256e-01	4.172924e-02
log_mgp_per_photo	1.292755e+00	2.531609e-01	1.431264e+00	2.951606e-01
log_step_pled_t	1.045138e-01	2.229695e-02	7.594344e-02	2.882616e-02
indie	2.157517e-01	6.061069e-02	2.079992e-01	6.114411e-02
is_site	4.603345e-01	6.876874e-02	4.567343e-01	6.939891e-02
mmo	-3.730230e-01	1.071894e-01	-3.767239e-01	1.081343e-01
log_goal	2.633207e-01	2.252159e-02	2.540411e-01	2.275840e-02
const	-5.255631e-01	1.847805e-01	-3.230511e-01	2.256832e-01
fighting	-2.489962e-01	7.705722e-02	-2.545736e-01	7.726201e-02
backed	8.577824e-03	3.305294e-03	8.441692e-03	3.431628e-03
date_year_2013	-3.267191e-01	7.984620e-02	-3.302566e-01	7.996221e-02
adventure	2.182955e-01	5.865141e-02	2.246345e-01	5.980526e-02
n_pled_t	9.157724e-02	1.167488e-02	8.801533e-02	1.220501e-02
rpg	1.575587e-01	6.376203e-02	1.656458e-01	6.490885e-02
text_am_sq	-2.292496e-07	2.689099e-08	-2.316835e-07	2.731744e-08
action	-1.168190e-01	6.110808e-02	-1.308194e-01	6.171504e-02

Figure 3: Comparing coefficients of final preliminary model and consistent model

On the figure above final_model is the final preliminary model and raw_model is the consistent model. As we can see, none of the coefficients differ by more than 1 std error, so I suppose that our final preliminary model is consistent as well.

7.2 Checking for the heteroscedasticity

Let's conduct smart White test for heteroscedasticity. I will create squares and cross-products of non-dummy variables only (still, simple dummies are not deleted from White test equation), since otherwise, there will be too much variables and also I think that at the same time these two event can happen extremely rarely:

1) cross-product of dummy with other variable affects standard error of disturbance term

2) Neither dummy variable nor any first or second order combination of continitous variables do not affect standard error of disturbance term (i.e. If continious variable * dummy affect std of dist term then just this continious variable is likely to affect std of dist term as well!!!)

So let's do this test:

```
y = final_model.resid ** 2
white_test_regression_final_model = sm.OLS(y, df_extended).fit(cov_type = 'HC3')
RSS_white_test_regression_final_model = RSS(y, white_test_regression_final_model)
print('RSS: {}'.format(RSS_white_test_regression_final_model))
display(white_test_regression_final_model.summary())
```

```
RSS: 178835.31231320451
```

```
D:\Anaconda\lib\site-packages\statsmodels\base\model.py:1832: ValueWarning: covariance of co
The number of constraints is 85, but rank is 74
  warnings.warn('covariance of constraints does not have full '
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.134
Model:	OLS	**Adj. R-squared:**	0.120
Method:	Least Squares	**F-statistic:**	3.215
Date:	Sat, 30 Apr 2022	**Prob (F-statistic):**	1.53e-18
Time:	12:46:11	**Log-Likelihood:**	-17056.
No. Observations:	5379	**AIC:**	3.428e+04
Df Residuals:	5293	**BIC:**	3.485e+04
Df Model:	85		
Covariance Type:	HC3		

Figure 4: Smart White test for the final preliminary model

As can be seen from the short output (long output can be observed in the Averin_applied_essay_econometrics_gr1_code.ipynb file in the zip file), both chi2-statistic and F-statistic imply that there is the presence of Heteroscedasticity.

In the Averin_applied_essay_econometrics_gr1_code.ipynb file in the zip file you can find my attempts to use WLS to eliminate heteroscedasticity, but they were unsuccessful, so I was unable to make my estimates more efficient, but still I can do significance test, because robust standard errors were used! So my final preliminary model becomes final! And I will proceed with this model further on:

$log\_pled_i = \beta_0 + \beta_1 text\_am_i + \beta_2 log\_n\_img_i + \beta_3 log\_mgp\_per\_photo_i + \beta_4 is\_site_i + \beta_5 mmo_i + \beta_6 backed_i + \beta_7 adventure_i + \beta_8 n\_pled\_t_i + \beta_9 text\_am\_sq_i + \beta_{10} log\_min\_pled\_t_i + \beta_{11} shooter_i + \beta_{12} n\_vid\_binary_i + \beta_{13} log\_cr\_time_i + \beta_{14} log\_step\_pled\_t_i + \beta_{15} indie_i + \beta_{16} fighting_i + \beta_{17} rpg_i + \beta_{18} log\_goal_i + \beta_{19} action_i + \beta_{20} date\_year\_2013_i + \beta_{21} date\_year\_2014\_15\_16_i + v_i$

8 Interpreting the model and estimation of the optimal number of words

8.1 Interpretation of the final model

So, our final model is:

$log\_pled_i = \beta_0 + \beta_1 text\_am_i + \beta_2 log\_n\_img_i + \beta_3 log\_mgp\_per\_photo_i + \beta_4 is\_site_i + \beta_5 mmo_i + \beta_6 backed_i + \beta_7 adventure_i + \beta_8 n\_pled\_t_i + \beta_9 text\_am\_sq_i + \beta_{10} log\_min\_pled\_t_i + \beta_{11} shooter_i + \beta_{12} n\_vid\_binary_i + \beta_{13} log\_cr\_time_i + \beta_{14} log\_step\_pled\_t_i + \beta_{15} indie_i + \beta_{16} fighting_i + \beta_{17} rpg_i + \beta_{18} log\_goal_i + \beta_{19} action_i + \beta_{20} date\_year\_2013_i + \beta_{21} date\_year\_2014\_15\_16_i + v_i$

And the estimation of it:

Dep. Variable:	log_pled	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	342.6
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:09	Log-Likelihood:	-11108.
No. Observations:	5379	AIC:	2.226e+04
Df Residuals:	5357	BIC:	2.240e+04
Df Model:	21		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
date_year_2014_15_16	-0.7172	0.072	-9.930	0.000	-0.859	-0.576
log_min_pled_t	0.1164	0.037	3.138	0.002	0.044	0.189
shooter	-0.1867	0.091	-2.062	0.039	-0.364	-0.009
n_vid_binary	1.6705	0.064	25.986	0.000	1.545	1.797
text_am	0.0013	0.000	10.487	0.000	0.001	0.002
log_cr_time	0.0691	0.016	4.365	0.000	0.038	0.100
log_n_img	0.4581	0.036	12.748	0.000	0.388	0.529
log_mgp_per_photo	1.2928	0.253	5.106	0.000	0.797	1.789
log_step_pled_t	0.1045	0.022	4.687	0.000	0.061	0.148
indie	0.2158	0.061	3.560	0.000	0.097	0.335
is_site	0.4603	0.069	6.694	0.000	0.326	0.595
mmo	-0.3730	0.107	-3.480	0.001	-0.583	-0.163
log_goal	0.2633	0.023	11.692	0.000	0.219	0.307
const	-0.5256	0.185	-2.844	0.004	-0.888	-0.163
fighting	-0.2490	0.077	-3.231	0.001	-0.400	-0.098
backed	0.0086	0.003	2.595	0.009	0.002	0.015
date_year_2013	-0.3267	0.080	-4.092	0.000	-0.483	-0.170
adventure	0.2183	0.059	3.722	0.000	0.103	0.333
n_pled_t	0.0916	0.012	7.844	0.000	0.069	0.114
rpg	0.1576	0.064	2.471	0.013	0.033	0.283
text_am_sq	-2.292e-07	2.69e-08	-8.525	0.000	-2.82e-07	-1.77e-07
action	-0.1168	0.061	-1.912	0.056	-0.237	0.003

Figure 5: Final model estimation

So let me interpret some of the parameters (not all, since there are to many of them):

1) log_min_pled_t: keeping other variables constant, increase in the minimum sum which can be donated to the project by 1%, increases the amount of funds raised by 0.1164% on average.

2) indie: keeping other variables constant, game project with indie tag in the description increases the amount of funds raised by 21.58% on average.

And most importantly:

3) text_am: if the project doesn't have any words in the description, keeping other variables constant, adding one word to the description increases the amount of funds raised by 0.13% on average.

4) text_am_sq: keeping other variables constant, with each additional word in the description, the marginal effect on pled of 1 added word to the description decreases on average by 0.00004584 percentage points.

Hooray!!! coefficient to text_am_sq is significantly less than zero, so we can estimate the optimal number of words for the gaming project description on Kickstarter.com

8.2 Estimation of the optimal number of words

To calculate it, let's present our equation in the following form:

$\widehat{log\_pled_i} = f(\widehat{control\_vars_i}) + \hat{\gamma_1} text\_am_i + \hat{\gamma_2} text\_am\_sq_i$

Thus, using simple school maths and the fact that $\hat{\gamma_2}$ is negative, the optimal text_am can be calculated as follows: optimal number of words in the description of the project $= -\frac{\hat{\gamma_1}}{2\hat{\gamma_2}} = -\frac{0.0013}{-2*10^{-7}*2.292} = 2772.88$

So the estimation of the optimal number of words in the game project description on Kickstarter.com is 2773!!!!

This is the sample estimation, so let us find out, what is the standard error of this parameter in the next section.

9 Sample distribution of the optimal number of words using Bootstrap method, calculating confidence intervals for this parameter

Since the optimal number of words is non-linear function of parameters of regression, (parameter to text_am_sq $\hat{\gamma_2}$ is in denominator), then it's complicated to find variance of the optimal number of words in theoretical way, so let me calculate confidence interval practically using Bootstrap method!!! This method is similar to Monte-Carlo, but instead of random generation of data, we resample from the existing data, by randomly choosing rows of the data WITH REPLACEMENT!!! I will generate 10000 bootstraped samples and will calculate optimal number of words for each sample, so for 95% confidence interval I will take 2.5 and 97.5 percentiles from the data of 10000 bootstrap estimates of the optimal word number!!! The sample is large, so bootstrap method should provide asymptotically valid confidence intervals!!

So the sample distribution of the optimal number of words using Bootstrap method:

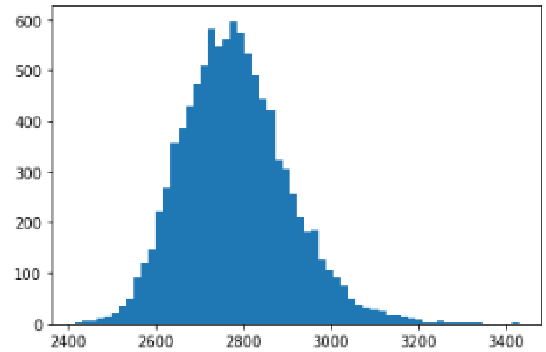

Figure 6: sample distribution of the optimal number of words using Bootstrap method

So we can see that the distribution is bell-shaped, but skewed, which makes it different from the normal distribution. As for me, it looks like scaled chi-2 distribution with large degrees of freedom.

Some statistics about optimal word number (using Bootstrap method):

```
Mean: 2782.184, Std: 121.084
95% confidence interval: [2570.092, 3041.317]
99% confidence interval: [2515.104, 3159.68]
```

Figure 7: mean, std, confidence intervals of the optimal number of words

Note that even 99% confidence interval above is fully within the range of the text_am in the sample, which makes our estimation and conclusion feasible to reach.

10 Conclusion

From the analysis of the game projects on Kickstarter.com from 2012-2016 (which raised funds in USD) it was empirically found out that that there exists the feasible optimal number of words in the description which given other factors constant, maximizes the amount of funds raised. The estimate of this number is 2770-2785 words with 95% confidence interval [2570, 3042] words. So game project developers can observe such optimal number of words and make the descriptions more revenue-generating. (Hope that this optimal number does not extrapolate to applied econometric essays since I wrote more that optimal number of words:))

11 Further analysis

1) Success of the project may be defined in the different way. Alternatively, succsess variable could have used, and logit estimation with the similar explanatory variables could have been done to find out optimal number of words. Similar if using the percent of funds raised relative to the target (goal)

2) optimal number of words may depend on genre of the game or have some seasonality, so it may be reasonable to consider some separate regressions (the amount of data allows to split the data and made further analysis like that)

3) Optimal parameters of other variables may be calculated if possible

4) text_am may not be quite deterministic: game developers may have some belief about the project success which correlates with real success of the game, and if they don't believe in the project, then they may put

less words in the description, and thus the effect of text_am in the current analysis is exaggerated, so it is possible here to think about some possible instrumental variables estimation.

5) Using more variables in the model (large sample allows to do so), i.e. cross-multiplication variables (slope dummies, and other cross-multiplications), different polynomials of the variables and some macro factors.

6) Projects which raised funds in other currencies rather than USD can be considered in further analysis as well.

7) The risk that analysis may not be extrapolated to current and future years: although quite robust results for 2014-2016 - last 3 years of the data.

8) Use more complicated approaches in finding out of the best model (i.e. run all the regressions with all the possible groups of the variables, which have alternative counterparts and so on)

12 Bibliography

1) Zamkov Oleg Olegovich's lecture slides.

13 Appendix

13.1 Description of variables in the initial data

URL: Link to the Kickstarter page of the game

pled: amount of money raised by the project, in USD

goal: amount of money which game developers wanted to raise to realise the project, in USD

date: date of publication of the project on Kickstarter.com

period: duration of the fund-raising, for how long inverstors were able to fund the project since its publication of Kickstarter, in days

Status: whether the project was still raising the funds or not as at the moment of parsing this data

text_am: number of words in the description of the project on Kickstarter

n_vid and n_img: number of videos and images in the description of the project in Kickstarter

game categories (rpg, platformer, shooter, fighting, survival, horror, strategy, arcade, simulator, mmo, indie, action, quest, adventure, mgp): 1 if such tag of category was mentioned in the description. Note, that one project may have several tags or none of them.

n_pled_t: number of the options (in terms of amount of money) for investor to fund the project, in USD

min_pled_t, max_pled_t, mean_pled_t: minimum/max/mean of donation options (in terms of amount of money), in USD

step_pled_t: the average difference between donation options, in USD

cr_time: time from the creator's account registration, in days

backed: the number of other projects, financed by the creator of the game projects

created: the number of the other projects published on Kickstarter by the creator before publication of this game (on Kickstarter)

mgp: sum of photo pixels in the description of the projects, in millions of pixels

cont: average contrast ration of the photos, in scaled unit where 0 - no contrast, 1 - maximum constrast

curr: currency, in which the project raises funds

succsess: whether the project funded more than planned (1 if more, 0 if less) site: separate site of the gaming project (if exists)

is_site: whether the project has its own site. (1 if yes, 0 if no)

13.2 The reason I added some more variables

1) 1st set: why I added some log variables:

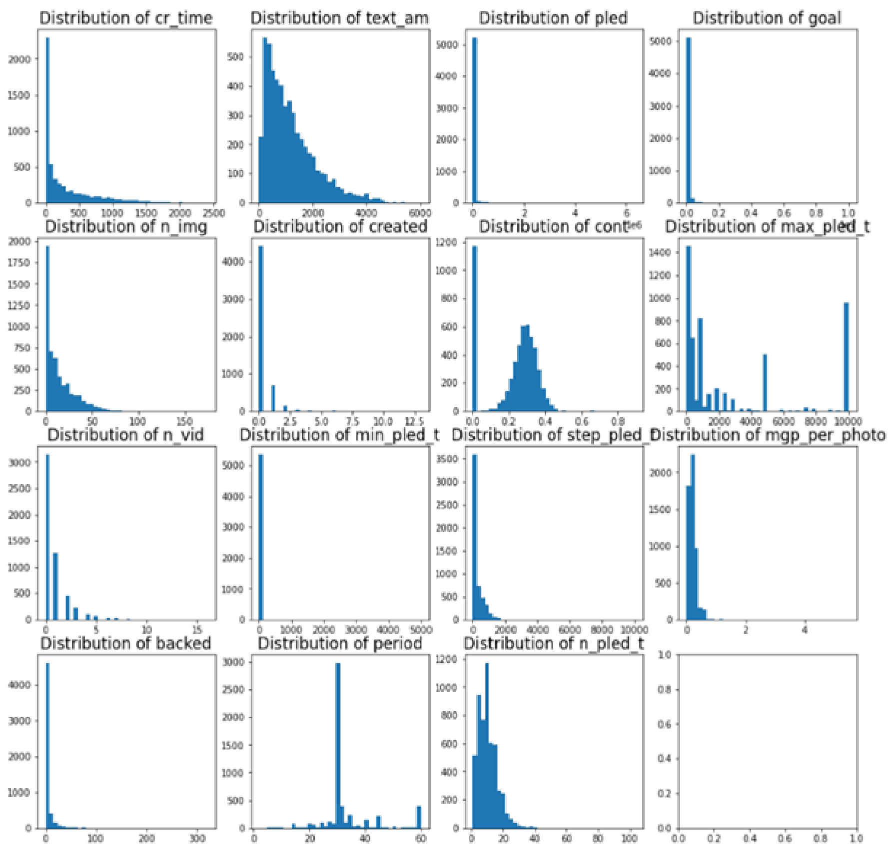

Figure 8: Distribution of initial numeric variables

From figure 4 it can be seen that variables goal, mgp_per_photo, n_img, cr_time, min_pled_t, step_pled_t and pled (pled was already discussed as it's a target variable) are very skewed to the right, but OLS fits better if the variable is closer to the normal distribution, so that is why I decided to take logarithm of the variables.

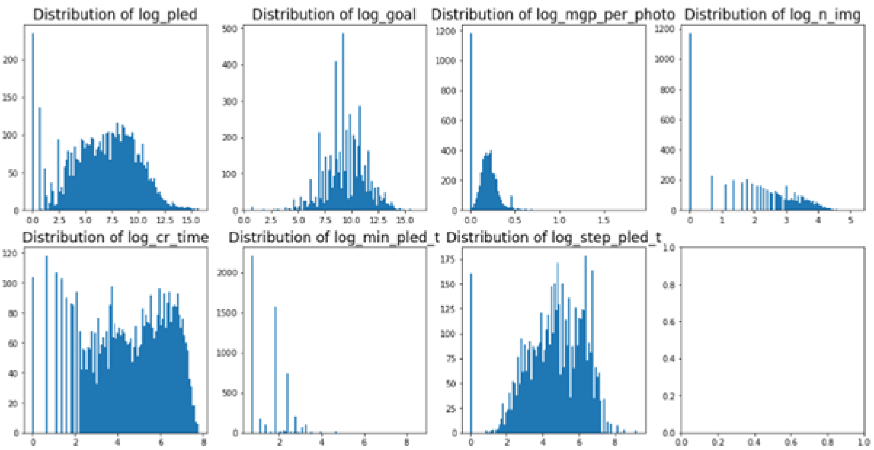

Figure 9: Distribution of the logarithms of the variables which were skewed to the left

From figure 5 it can be seen that the logarithms of these variables is less skewed and in some cases it becomes very close to the normal distribution, so this may improve the model

2) Why I have created binary variables, related to n_img and n_vid:

Here, I just made the assumption that the presence of the video/photo may be an important factor and even more important factor than the number of photos/videos in the description.

3) Why I have created additional tags related to the tags:

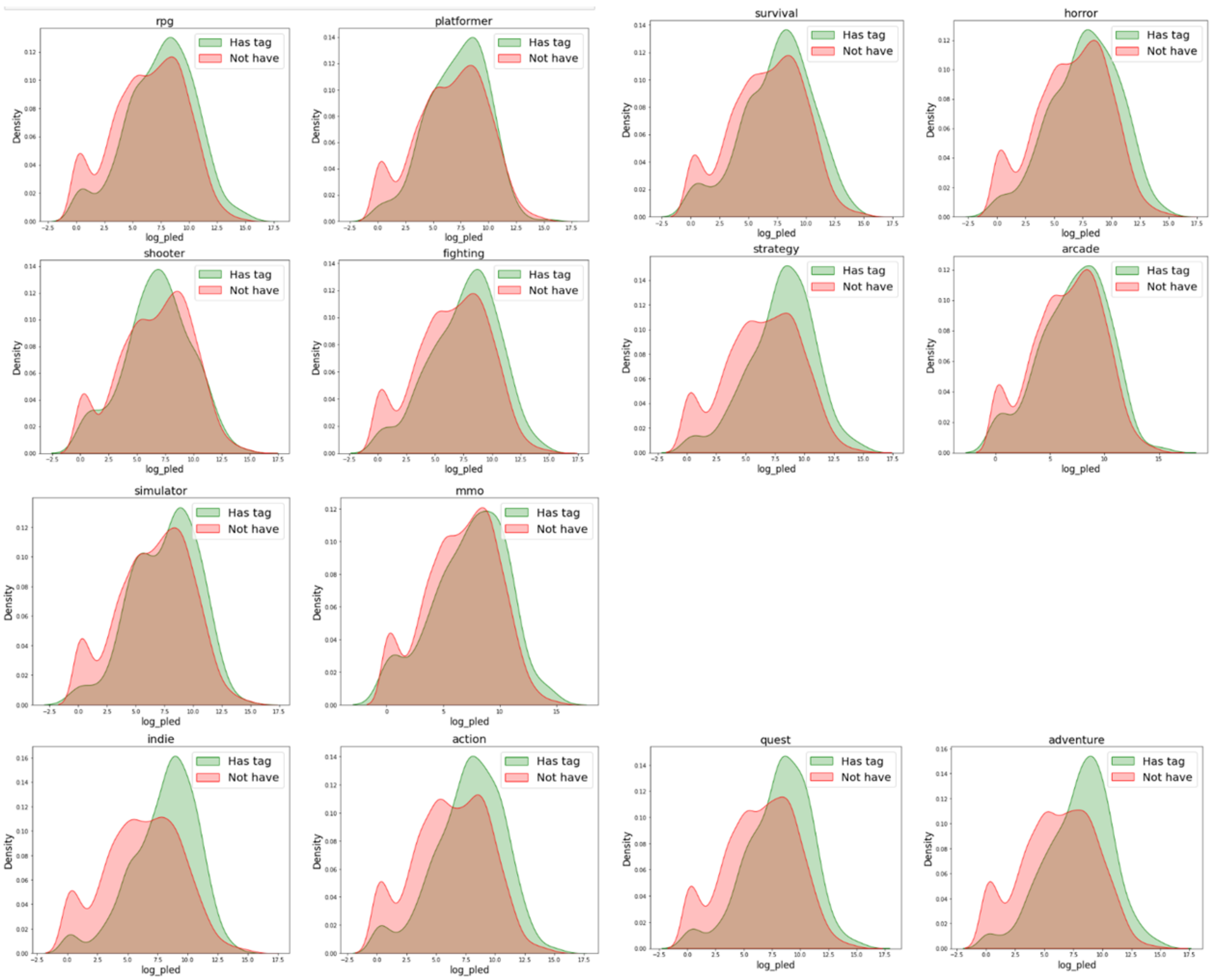

Figure 10: The distribution of the log_pled for projects which have the certain tag and which does not have the certain tag

We can see for some tags the presence of the tag increases log_pled on average, but maybe the number of the tags is more important than which tags are present?

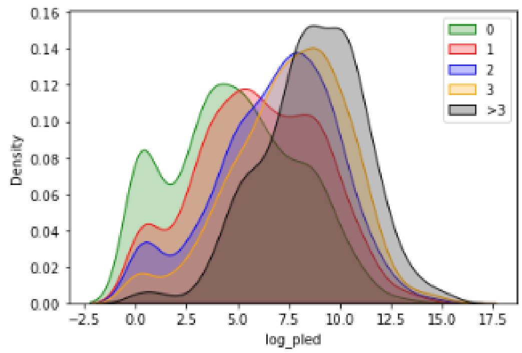

Figure 11: The distribution of the log_pled for projects with certain number of tags

The distribution is different for different number of tags, so maybe these dummy variables (related to the number of tags) should be used instead of the initial dummy tag variables

4) Why I have created variables related to date:

I found it reasonable to include the time in the model, as there may be a time trend: as the time moves on there may be more and more investors on Kickstarter.com, so on average the game project is going to attract more and more funds. However, the may be seasonality, so in some particular year Kickstarter.com was popular than in some other years. That is why I created the dummy year variables.

13.3 Variables selection

a) Set 1: (goal, mgp_per_photo, n_img, cr_time, min_pled_t, step_pled_t) VS (log_goal, log_mgp_per_photo, log_n_img, log_cr_time, log_min_pled_t, log_step_pled_t): it is better to use log variables

Dep. Variable:	log_pled	R-squared:	0.592
Model:	OLS	Adj. R-squared:	0.590
Method:	Least Squares	F-statistic:	166.2
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:07	Log-Likelihood:	-11457.
No. Observations:	5379	AIC:	2.298e+04
Df Residuals:	5347	BIC:	2.319e+04
Df Model:	31		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
const	13.5808	1.093	12.428	0.000	11.439	15.723
period	0.0027	0.003	0.890	0.373	-0.003	0.009
text_am	0.0019	0.000	14.032	0.000	0.002	0.002
n_pled_t	0.0949	0.016	6.104	0.000	0.064	0.125
max_pled_t	8.846e-05	2.69e-05	3.290	0.001	3.58e-05	0.000
backed	0.0066	0.004	1.740	0.082	-0.001	0.014
created	-0.0363	0.037	-0.986	0.324	-0.109	0.036
cont	1.8728	0.373	5.024	0.000	1.142	2.603
is_site	0.5964	0.072	8.243	0.000	0.455	0.738
text_am_sq	-3.733e-07	3.16e-08	-11.810	0.000	-4.35e-07	-3.11e-07
goal	4.116e-07	3.36e-07	1.226	0.220	-2.46e-07	1.07e-06
mgp_per_photo	1.0715	0.407	2.633	0.008	0.274	1.869
n_img	0.0335	0.003	10.704	0.000	0.027	0.040
cr_time	0.0005	7.95e-05	6.126	0.000	0.000	0.001
min_pled_t	-0.0003	0.001	-0.279	0.780	-0.003	0.002
step_pled_t	-8.905e-05	0.000	-0.387	0.699	-0.001	0.000
n_vid	0.3133	0.027	11.425	0.000	0.260	0.367
rpg	0.1662	0.069	2.418	0.016	0.032	0.301
platformer	0.0108	0.098	0.111	0.912	-0.181	0.202
shooter	-0.1721	0.096	-1.783	0.075	-0.361	0.017
fighting	-0.2793	0.083	-3.372	0.001	-0.442	-0.117
survival	0.0277	0.086	0.321	0.748	-0.141	0.197
horror	0.1152	0.099	1.162	0.245	-0.079	0.309
strategy	0.0412	0.076	0.540	0.589	-0.108	0.191
arcade	0.1743	0.097	1.796	0.072	-0.016	0.365
simulator	0.3128	0.149	2.099	0.036	0.021	0.605
mmo	-0.3733	0.112	-3.342	0.001	-0.592	-0.154
indie	0.2141	0.067	3.186	0.001	0.082	0.346
action	-0.1341	0.065	-2.053	0.040	-0.262	-0.006
quest	0.0472	0.084	0.564	0.573	-0.117	0.211
adventure	0.2926	0.065	4.481	0.000	0.165	0.421
date_timestamp	-8.425e-09	7.56e-10	-11.151	0.000	-9.91e-09	-6.94e-09

Omnibus:	291.855	Durbin-Watson:	1.963
Prob(Omnibus):	0.000	Jarque-Bera (JB):	396.583
Skew:	-0.507	Prob(JB):	7.64e-87
Kurtosis:	3.861	Cond. No.	5.06e+10

Dep. Variable:	log_pled	R-squared:	0.611
Model:	OLS	Adj. R-squared:	0.609
Method:	Least Squares	F-statistic:	189.9
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:07	Log-Likelihood:	-11331.
No. Observations:	5379	AIC:	2.273e+04
Df Residuals:	5347	BIC:	2.294e+04
Df Model:	31		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
text_am	0.0015	0.000	11.444	0.000	0.001	0.002
created	-0.0160	0.035	-0.454	0.650	-0.085	0.053
quest	0.0565	0.081	0.702	0.483	-0.101	0.214
max_pled_t	1.717e-05	1.31e-05	1.312	0.189	-8.47e-06	4.28e-05
is_site	0.5382	0.072	7.468	0.000	0.397	0.679
mmo	-0.3818	0.110	-3.462	0.000	-0.597	-0.167
n_vid	0.3150	0.027	11.651	0.000	0.262	0.368
backed	0.0088	0.004	2.254	0.024	0.001	0.016
strategy	0.0007	0.074	0.009	0.993	-0.145	0.146
simulator	0.2640	0.150	1.757	0.079	-0.030	0.558
period	-0.0008	0.003	-0.279	0.780	-0.007	0.005
adventure	0.2364	0.063	3.745	0.000	0.113	0.360
n_pled_t	0.0946	0.013	7.394	0.000	0.070	0.120
text_am_sq	-2.806e-07	2.86e-08	-9.804	0.000	-3.37e-07	-2.25e-07
shooter	-0.1955	0.095	-2.065	0.039	-0.381	-0.010
horror	0.1014	0.097	1.044	0.296	-0.089	0.292
arcade	0.1901	0.094	2.025	0.043	0.006	0.374
survival	-0.0173	0.083	-0.208	0.836	-0.181	0.146
platformer	0.0371	0.096	0.387	0.699	-0.151	0.225
date_timestamp	-7.053e-09	7.03e-10	-10.027	0.000	-8.43e-09	-5.67e-09
cont	-0.4870	0.355	-1.373	0.170	-1.182	0.208
indie	0.2402	0.065	3.699	0.000	0.113	0.367
const	9.4677	1.030	9.192	0.000	7.449	11.486
fighting	-0.2708	0.081	-3.338	0.001	-0.430	-0.112
rpg	0.2142	0.068	3.155	0.002	0.081	0.347
action	-0.1677	0.064	-2.620	0.009	-0.293	-0.042
log_goal	0.1938	0.023	8.501	0.000	0.149	0.239
log_mgp_per_photo	2.0166	0.330	6.104	0.000	1.369	2.664
log_n_img	0.5978	0.043	13.748	0.000	0.513	0.683
log_cr_time	0.0836	0.017	5.005	0.000	0.051	0.116
log_min_pled_t	0.1314	0.039	3.379	0.001	0.055	0.208
log_step_pled_t	0.0748	0.029	2.557	0.011	0.017	0.132

Omnibus:	347.447	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	470.289
Skew:	-0.578	Prob(JB):	7.55e-103
Kurtosis:	3.874	Cond. No.	5.04e+10

Figure 12: Regression without logs VS regression with logs

```
chi2_st(1) = 126.56309896501728
chi2_crit(1%, df = 1) = 6.6348966010212145
```

Figure 13: Zarembka test

As can be seen from the regression outputs, regression with logarithms has lower information criterion and Zarembka test shows that RSS of regression with logarithms is significantly lower. Thus, choosing log variables further on.

b) Set 2a: (n_vid_binary) VS (n_vid): it is better to use binary variable

Dep. Variable:	log_pled	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	226.8
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:07	Log-Likelihood:	-11106.
No. Observations:	5379	AIC:	2.228e+04
Df Residuals:	5347	BIC:	2.249e+04
Df Model:	31		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
text_am	0.0013	0.000	10.200	0.000	0.001	0.001
log_n_img	0.4762	0.041	11.524	0.000	0.395	0.557
created	-0.0036	0.033	-0.109	0.914	-0.069	0.061
log_mgp_per_photo	1.4622	0.296	4.932	0.000	0.881	2.043
quest	0.0560	0.076	0.733	0.463	-0.094	0.206
max_pled_t	2.11e-05	1.26e-05	1.673	0.094	-3.62e-06	4.58e-05
is_site	0.4592	0.070	6.593	0.000	0.323	0.596
mmo	-0.3800	0.107	-3.537	0.000	-0.591	-0.169
backed	0.0087	0.003	2.613	0.009	0.002	0.015
strategy	0.0707	0.071	0.995	0.320	-0.069	0.210
simulator	0.2397	0.142	1.683	0.092	-0.039	0.519
period	5.923e-05	0.003	0.020	0.984	-0.006	0.006
adventure	0.2161	0.060	3.617	0.000	0.099	0.333
n_pled_t	0.0872	0.012	7.157	0.000	0.063	0.111
text_am_sq	-2.324e-07	2.74e-08	-8.481	0.000	-2.86e-07	-1.79e-07
log_min_pled_t	0.1100	0.038	2.927	0.003	0.036	0.184
shooter	-0.1734	0.092	-1.880	0.060	-0.354	0.007
horror	0.1038	0.090	1.148	0.251	-0.073	0.281
arcade	0.1558	0.089	1.748	0.080	-0.019	0.330
survival	-0.0464	0.080	-0.579	0.563	-0.204	0.111
platformer	0.0397	0.089	0.446	0.655	-0.135	0.214
log_cr_time	0.0667	0.016	4.125	0.000	0.035	0.098
date_timestamp	-6.668e-09	6.71e-10	-9.944	0.000	-7.98e-09	-5.35e-09
log_step_pled_t	0.0760	0.029	2.634	0.008	0.019	0.133
cont	-0.3444	0.337	-1.022	0.307	-1.005	0.316
indie	0.1869	0.061	3.078	0.002	0.068	0.306
const	8.5404	0.984	8.677	0.000	6.611	10.470
fighting	-0.2542	0.077	-3.294	0.001	-0.405	-0.103
rpg	0.1781	0.065	2.738	0.006	0.051	0.306
log_goal	0.2565	0.023	11.242	0.000	0.212	0.301
action	-0.1248	0.062	-2.028	0.043	-0.245	-0.004
n_vid_binary	1.6632	0.064	25.877	0.000	1.537	1.789

Omnibus:	351.882	Durbin-Watson:	1.994
Prob(Omnibus):	0.000	Jarque-Bera (JB):	476.020
Skew:	-0.583	Prob(JB):	4.30e-104
Kurtosis:	3.874	Cond. No.	5.04e+10

Dep. Variable:	log_pled	R-squared:	0.611
Model:	OLS	Adj. R-squared:	0.609
Method:	Least Squares	F-statistic:	189.9
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:07	Log-Likelihood:	-11331.
No. Observations:	5379	AIC:	2.273e+04
Df Residuals:	5347	BIC:	2.294e+04
Df Model:	31		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
text_am	0.0015	0.000	11.444	0.000	0.001	0.002
created	-0.0160	0.035	-0.454	0.650	-0.085	0.053
quest	0.0565	0.081	0.702	0.483	-0.101	0.214
max_pled_t	1.717e-05	1.31e-05	1.312	0.189	-8.47e-06	4.28e-05
is_site	0.5382	0.072	7.468	0.000	0.397	0.679
mmo	-0.3818	0.110	-3.482	0.000	-0.597	-0.167
n_vid	0.3150	0.027	11.651	0.000	0.262	0.368
backed	0.0088	0.004	2.254	0.024	0.001	0.016
strategy	0.0007	0.074	0.009	0.993	-0.145	0.146
simulator	0.2640	0.150	1.757	0.079	-0.030	0.558
period	-0.0008	0.003	-0.279	0.780	-0.007	0.005
adventure	0.2364	0.063	3.745	0.000	0.113	0.360
n_pled_t	0.0946	0.013	7.394	0.000	0.070	0.120
text_am_sq	-2.806e-07	2.86e-08	-9.804	0.000	-3.37e-07	-2.25e-07
shooter	-0.1955	0.095	-2.065	0.039	-0.381	-0.010
horror	0.1014	0.097	1.044	0.296	-0.089	0.292
arcade	0.1901	0.094	2.025	0.043	0.006	0.374
survival	-0.0173	0.083	-0.208	0.836	-0.181	0.146
platformer	0.0371	0.096	0.387	0.699	-0.151	0.225
date_timestamp	-7.053e-09	7.03e-10	-10.027	0.000	-8.43e-09	-5.67e-09
cont	-0.4870	0.355	-1.373	0.170	-1.182	0.208
indie	0.2402	0.065	3.699	0.000	0.113	0.367
const	9.4677	1.030	9.192	0.000	7.449	11.486
fighting	-0.2708	0.081	-3.338	0.001	-0.430	-0.112
rpg	0.2142	0.068	3.155	0.002	0.081	0.347
action	-0.1677	0.064	-2.620	0.009	-0.293	-0.042
log_goal	0.1938	0.023	8.501	0.000	0.149	0.239
log_mgp_per_photo	2.0166	0.330	6.104	0.000	1.369	2.664
log_n_img	0.5978	0.043	13.748	0.000	0.513	0.683
log_cr_time	0.0836	0.017	5.005	0.000	0.051	0.116
log_min_pled_t	0.1314	0.039	3.379	0.001	0.055	0.208
log_step_pled_t	0.0748	0.029	2.557	0.011	0.017	0.132

Omnibus:	347.447	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	470.289
Skew:	-0.578	Prob(JB):	7.55e-103
Kurtosis:	3.874	Cond. No.	5.04e+10

Figure 14: Regression with n_vid_binary VS regression with n_vid

```
chi2_st(1) = 224.36196135731072
chi2_crit(1%, df = 1) = 6.6348966010212145
```

Figure 15: Zarembka test

As can be seen from the regression outputs, regression with n_vid_binary has lower information criterion and Zarembka test shows that RSS of regression with n_vid_binary is significantly lower. Thus, choosing n_vid_binary further on.

Set 2b: whether n_img_binary is significant: No, this variable is not significant

Dep. Variable:	log_pled	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	219.9
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:07	Log-Likelihood:	-11106.
No. Observations:	5379	AIC:	2.226e+04
Df Residuals:	5346	BIC:	2.250e+04
Df Model:	32		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
text_am	0.0013	0.000	10.188	0.000	0.001	0.001
log_n_img	0.4904	0.050	9.710	0.000	0.391	0.589
created	-0.0036	0.033	-0.110	0.913	-0.069	0.061
log_mgp_per_photo	1.5389	0.347	4.435	0.000	0.859	2.219
quest	0.0562	0.076	0.736	0.462	-0.094	0.206
max_pled_t	2.088e-05	1.26e-05	1.655	0.098	-3.84e-06	4.56e-05
is_site	0.4584	0.070	6.583	0.000	0.322	0.595
mmo	-0.3806	0.107	-3.542	0.000	-0.591	-0.170
backed	0.0087	0.003	2.611	0.009	0.002	0.015
strategy	0.0705	0.071	0.991	0.322	-0.069	0.210
simulator	0.2410	0.142	1.693	0.090	-0.038	0.520
period	2.169e-05	0.003	0.007	0.994	-0.006	0.006
adventure	0.2164	0.060	3.620	0.000	0.099	0.334
n_pled_t	0.0867	0.012	7.033	0.000	0.063	0.111
text_am_sq	-2.334e-07	2.76e-08	-8.463	0.000	-2.87e-07	-1.79e-07
log_min_pled_t	0.1087	0.038	2.879	0.004	0.035	0.183
shooter	-0.1709	0.092	-1.849	0.064	-0.352	0.010
horror	0.1047	0.090	1.158	0.247	-0.073	0.282
arcade	0.1554	0.089	1.744	0.081	-0.019	0.330
survival	-0.0461	0.080	-0.574	0.566	-0.203	0.111
platformer	0.0407	0.089	0.457	0.647	-0.134	0.215
log_cr_time	0.0667	0.016	4.128	0.000	0.035	0.098
date_timestamp	-6.73e-09	6.85e-10	-9.821	0.000	-8.07e-09	-5.39e-09
log_step_pled_t	0.0767	0.029	2.650	0.008	0.020	0.133
cont	-0.1740	0.425	-0.410	0.682	-1.007	0.659
indie	0.1865	0.061	3.070	0.002	0.067	0.306
const	8.6424	1.009	8.567	0.000	6.665	10.620
fighting	-0.2545	0.077	-3.297	0.001	-0.406	-0.103
rpg	0.1778	0.065	2.734	0.006	0.050	0.305
log_goal	0.2563	0.023	11.243	0.000	0.212	0.301
action	-0.1252	0.062	-2.033	0.042	-0.246	-0.005
n_vid_binary	1.6616	0.064	25.774	0.000	1.535	1.788
n_img_binary	-0.1134	0.207	-0.547	0.584	-0.520	0.293

Figure 16: Regression with n_img_binary

As can be seen from the p-value in the regression output above the coefficient to n_img_binary is not significant using t-test, so it is not used further on

c) Set 3: (rpg, platformer, shooter, fighting, survival, horror, strategy, arcade, simulator, mmo, indie, action, quest, adventure, mgp) VS (tags_sum) VS (tags_sum1_1, tags_sum1_2, tags_sum1_3, tags_sum1_¿3): better to use (rpg, platformer, shooter, fighting, survival, horror, strategy, arcade, simulator, mmo, indie, action, quest, adventure, mgp)

Dep. Variable:	log_pled	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	226.8
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:08	Log-Likelihood:	-11106.
No. Observations:	5379	AIC:	2.228e+04
Df Residuals:	5347	BIC:	2.249e+04
Df Model:	31		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
text_am	0.0013	0.000	10.200	0.000	0.001	0.001
log_n_img	0.4762	0.041	11.524	0.000	0.395	0.557
created	-0.0036	0.033	-0.109	0.914	-0.069	0.061
log_mgp_per_photo	1.4622	0.296	4.932	0.000	0.881	2.043
quest	0.0560	0.076	0.733	0.463	-0.094	0.206
max_pled_t	2.11e-05	1.26e-05	1.673	0.094	-3.62e-06	4.58e-05
is_site	0.4592	0.070	6.593	0.000	0.323	0.596
mmo	-0.3800	0.107	-3.537	0.000	-0.591	-0.169
backed	0.0087	0.003	2.613	0.009	0.002	0.015
strategy	0.0707	0.071	0.995	0.320	-0.069	0.210
simulator	0.2397	0.142	1.683	0.092	-0.039	0.519
period	5.923e-05	0.003	0.020	0.984	-0.006	0.006
adventure	0.2161	0.060	3.617	0.000	0.099	0.333
n_pled_t	0.0872	0.012	7.157	0.000	0.063	0.111
text_am_sq	-2.324e-07	2.74e-08	-8.481	0.000	-2.86e-07	-1.79e-07
log_min_pled_t	0.1100	0.038	2.927	0.003	0.036	0.184
shooter	-0.1734	0.092	-1.880	0.060	-0.354	0.007
horror	0.1038	0.090	1.148	0.251	-0.073	0.281
arcade	0.1558	0.089	1.748	0.080	-0.019	0.330
survival	-0.0464	0.080	-0.579	0.563	-0.204	0.111
platformer	0.0397	0.089	0.446	0.655	-0.135	0.214
log_cr_time	0.0667	0.016	4.125	0.000	0.035	0.098
date_timestamp	-6.668e-09	6.71e-10	-9.944	0.000	-7.98e-09	-5.35e-09
log_step_pled_t	0.0760	0.029	2.634	0.008	0.019	0.133
cont	-0.3444	0.337	-1.022	0.307	-1.005	0.316
indie	0.1869	0.061	3.078	0.002	0.068	0.306
const	8.5404	0.984	8.677	0.000	6.611	10.470
fighting	-0.2542	0.077	-3.294	0.001	-0.405	-0.103
rpg	0.1781	0.065	2.738	0.006	0.051	0.306
log_goal	0.2565	0.023	11.242	0.000	0.212	0.301
action	-0.1248	0.062	-2.028	0.043	-0.245	-0.004
n_vid_binary	1.6632	0.064	25.877	0.000	1.537	1.789

Dep. Variable:	log_pled	R-squared:	0.636
Model:	OLS	Adj. R-squared:	0.636
Method:	Least Squares	F-statistic:	389.7
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:08	Log-Likelihood:	-11139.
No. Observations:	5379	AIC:	2.232e+04
Df Residuals:	5360	BIC:	2.244e+04
Df Model:	18		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
log_min_pled_t	0.1105	0.038	2.934	0.003	0.037	0.184
backed	0.0099	0.003	2.900	0.004	0.003	0.017
n_vid_binary	1.6902	0.064	26.303	0.000	1.564	1.816
text_am	0.0013	0.000	10.224	0.000	0.001	0.002
log_cr_time	0.0686	0.016	4.238	0.000	0.037	0.100
log_n_img	0.4916	0.041	11.882	0.000	0.411	0.573
created	-0.0035	0.033	-0.104	0.917	-0.068	0.061
log_mgp_per_photo	1.4766	0.296	4.989	0.000	0.897	2.057
period	0.0002	0.003	0.059	0.953	-0.006	0.006
date_timestamp	-6.637e-09	6.69e-10	-9.918	0.000	-7.95e-09	-5.33e-09
log_step_pled_t	0.0791	0.029	2.736	0.006	0.022	0.136
cont	-0.4153	0.336	-1.237	0.216	-1.074	0.243
log_goal	0.2470	0.023	10.857	0.000	0.202	0.292
max_pled_t	1.721e-05	1.26e-05	1.367	0.172	-7.47e-06	4.19e-05
is_site	0.4582	0.069	6.599	0.000	0.322	0.594
n_pled_t	0.0875	0.012	7.078	0.000	0.063	0.112
text_am_sq	-2.363e-07	2.76e-08	-8.550	0.000	-2.91e-07	-1.82e-07
const	8.5592	0.983	8.703	0.000	6.632	10.487
tags_sum	0.0255	0.018	1.383	0.167	-0.011	0.062

Dep. Variable:	log_pled	R-squared:	0.638
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	329.0
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:08	Log-Likelihood:	-11136.
No. Observations:	5379	AIC:	2.232e+04
Df Residuals:	5357	BIC:	2.246e+04
Df Model:	21		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
log_min_pled_t	0.1107	0.038	2.936	0.003	0.037	0.185
backed	0.0099	0.003	2.880	0.004	0.003	0.017
n_vid_binary	1.6878	0.064	26.239	0.000	1.562	1.814
text_am	0.0012	0.000	10.034	0.000	0.001	0.001
log_cr_time	0.0686	0.016	4.238	0.000	0.037	0.100
log_n_img	0.4899	0.041	11.850	0.000	0.409	0.571
created	-0.0038	0.033	-0.115	0.908	-0.069	0.061
log_mgp_per_photo	1.4893	0.296	5.035	0.000	0.910	2.069
period	0.0002	0.003	0.075	0.940	-0.006	0.006
date_timestamp	-6.636e-09	6.7e-10	-9.900	0.000	-7.95e-09	-5.32e-09
log_step_pled_t	0.0780	0.029	2.703	0.007	0.021	0.135
cont	-0.4517	0.337	-1.342	0.180	-1.112	0.208
log_goal	0.2463	0.023	10.831	0.000	0.202	0.291
max_pled_t	1.784e-05	1.26e-05	1.417	0.156	-6.83e-06	4.25e-05
is_site	0.4598	0.069	6.618	0.000	0.324	0.596

I decided to leave initial dummy variables, because the tags (genres) are more important than the number of tags. Although BIC in the model with tags_sum is the lowest out of three candidates above, the variable tags_sum itself is insignificant, but several dummy variables of tags are significant, so optimal strategy here is to leave initial dummies and then delete some dummies which are not significant as a group!!!

d) Set 4: (date_timestamp) VS (date_year_2013, date_year_2014, date_year_2015, date_year_2016): better to use dummy variables.

Dep. Variable:	log_pled	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	226.8
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:08	Log-Likelihood:	-11106.
No. Observations:	5379	AIC:	2.226e+04
Df Residuals:	5347	BIC:	2.249e+04
Df Model:	31		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
text_am	0.0013	0.000	10.200	0.000	0.001	0.001
log_n_img	0.4782	0.041	11.524	0.000	0.395	0.557
created	-0.0036	0.033	-0.109	0.914	-0.069	0.061
log_mgp_per_photo	1.4622	0.296	4.932	0.000	0.881	2.043
quest	0.0560	0.076	0.733	0.463	-0.094	0.206
max_pled_t	2.11e-05	1.26e-05	1.673	0.094	-3.62e-06	4.58e-05
is_site	0.4592	0.070	6.593	0.000	0.323	0.596
mmo	-0.3800	0.107	-3.537	0.000	-0.591	-0.169
backed	0.0087	0.003	2.613	0.009	0.002	0.015
strategy	0.0707	0.071	0.995	0.320	-0.069	0.210
simulator	0.2397	0.142	1.683	0.092	-0.039	0.519
period	5.923e-05	0.003	0.020	0.984	-0.006	0.006
adventure	0.2161	0.060	3.617	0.000	0.099	0.333
n_pled_t	0.0872	0.012	7.157	0.000	0.063	0.111
text_am_sq	-2.324e-07	2.74e-08	-8.481	0.000	-2.86e-07	-1.79e-07
log_min_pled_t	0.1100	0.038	2.927	0.003	0.036	0.184
shooter	-0.1734	0.092	-1.880	0.060	-0.354	0.007
horror	0.1038	0.090	1.148	0.251	-0.073	0.281
arcade	0.1558	0.089	1.748	0.080	-0.019	0.330
survival	-0.0464	0.080	-0.579	0.563	-0.204	0.111
platformer	0.0397	0.089	0.446	0.655	-0.135	0.214
log_cr_time	0.0667	0.016	4.125	0.000	0.035	0.098
date_timestamp	-6.668e-09	6.71e-10	-9.944	0.000	-7.98e-09	-5.35e-09
log_step_pled_t	0.0760	0.029	2.634	0.008	0.019	0.133
cont	-0.3444	0.337	-1.022	0.307	-1.005	0.316
indie	0.1869	0.061	3.078	0.002	0.068	0.306
const	8.5404	0.984	8.677	0.000	6.611	10.470
fighting	-0.2542	0.077	-3.294	0.001	-0.405	-0.103
rpg	0.1781	0.065	2.738	0.006	0.051	0.306
log_goal	0.2565	0.023	11.242	0.000	0.212	0.301
action	-0.1248	0.062	-2.028	0.043	-0.245	-0.004
n_vid_binary	1.6632	0.064	25.877	0.000	1.537	1.789

Dep. Variable:	log_pled	R-squared:	0.643
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	209.4
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:08	Log-Likelihood:	-11101.
No. Observations:	5379	AIC:	2.227e+04
Df Residuals:	5344	BIC:	2.250e+04
Df Model:	34		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
text_am	0.0013	0.000	10.209	0.000	0.001	0.001
log_n_img	0.4862	0.042	11.652	0.000	0.404	0.568
created	-0.0086	0.033	-0.262	0.793	-0.073	0.056
log_mgp_per_photo	1.4313	0.295	4.849	0.000	0.853	2.010
quest	0.0562	0.077	0.733	0.463	-0.094	0.206
max_pled_t	2.02e-05	1.26e-05	1.604	0.109	-4.48e-06	4.49e-05
is_site	0.4567	0.069	6.581	0.000	0.321	0.593
mmo	-0.3767	0.108	-3.484	0.000	-0.589	-0.165
backed	0.0084	0.003	2.460	0.014	0.002	0.015
strategy	0.0864	0.071	1.216	0.224	-0.053	0.226
simulator	0.2383	0.143	1.665	0.096	-0.042	0.519
period	0.0002	0.003	0.057	0.955	-0.006	0.006
adventure	0.2246	0.060	3.756	0.000	0.107	0.342
n_pled_t	0.0880	0.012	7.211	0.000	0.064	0.112
text_am_sq	-2.317e-07	2.73e-08	-8.481	0.000	-2.85e-07	-1.78e-07
log_min_pled_t	0.1095	0.038	2.917	0.004	0.036	0.183
shooter	-0.1867	0.092	-2.035	0.042	-0.366	-0.007
n_vid_binary	1.6652	0.064	25.818	0.000	1.539	1.792
horror	0.0968	0.090	1.079	0.281	-0.079	0.273
arcade	0.1649	0.090	1.841	0.066	-0.011	0.340
survival	-0.0380	0.080	-0.475	0.635	-0.195	0.119
platformer	0.0371	0.089	0.417	0.676	-0.137	0.211
log_cr_time	0.0711	0.016	4.394	0.000	0.039	0.103
log_step_pled_t	0.0759	0.029	2.635	0.008	0.019	0.132
cont	-0.3987	0.337	-1.182	0.237	-1.060	0.263
indie	0.2080	0.061	3.402	0.001	0.088	0.328
const	-0.3231	0.226	-1.431	0.152	-0.765	0.119
fighting	-0.2546	0.077	-3.295	0.001	-0.406	-0.103
rpg	0.1656	0.065	2.552	0.011	0.038	0.293
log_goal	0.2540	0.023	11.163	0.000	0.209	0.299
action	-0.1308	0.062	-2.120	0.034	-0.252	-0.010
date_year_2013	-0.3303	0.080	-4.130	0.000	-0.487	-0.174
date_year_2014	-0.7115	0.088	-8.096	0.000	-0.884	-0.539
date_year_2015	-0.7791	0.086	-9.010	0.000	-0.949	-0.610
date_year_2016	-0.7171	0.095	-7.555	0.000	-0.903	-0.531

Figure 18: Regression with date_timestamp VS Regression with date year dummies

If dummy then AIC stays the same, but BIC increases. Still I proceed with dummies, since it can be seen that in 2012 and 2013 log_pled was significantly lower on average (given other variables equal). Also estimates for 2014, 2015, 2016 do not differ significantly from each other, so I am going to try to aggregate these variables further on which may increase AIC and BIC.

13.4 Deleting/aggregating variables

After choosing the best variables the equation is as follows:

$log\_pled_i = \beta_0 + \beta_1 text\_am_i + \beta_2 log\_n\_img_i + \beta_3 created_i + \beta_4 log\_mgp\_per\_photo_i + \beta_5 quest_i + \beta_6 max\_pled\_t_i + \beta_7 is\_site_i + \beta_8 mmo_i + \beta_9 backed_i + \beta_{10} strategy_i + \beta_{11} simulator_i + \beta_{12} period_i + \beta_{13} adventure_i + \beta_{14} n\_pled\_t_i + \beta_{15} text\_am\_sq_i + \beta_{16} log\_min\_pled\_t_i + \beta_{17} shooter_i + \beta_{18} n\_vid\_binary_i + \beta_{19} horror_i + \beta_{20} arcade_i + \beta_{21} survival_i + \beta_{22} platformer_i + \beta_{23} log\_cr\_time_i + \beta_{24} log\_step\_pled\_t_i + \beta_{25} cont_i + \beta_{26} indie_i + \beta_{27} fighting_i + \beta_{28} rpg_i + \beta_{29} log\_goal_i + \beta_{30} action_i + \beta_{31} date\_year\_2013_i + \beta_{32} date\_year\_2014_i + \beta_{33} date\_year\_2015_i + \beta_{34} date\_year\_2016_i + v_i$

And regression output for the equation above is as follows:

Dep. Variable:	log_pled	**R-squared:**	0.643
Model:	OLS	**Adj. R-squared:**	0.640
Method:	Least Squares	**F-statistic:**	209.4
Date:	Sat, 30 Apr 2022	**Prob (F-statistic):**	0.00
Time:	12:46:09	**Log-Likelihood:**	-11101.
No. Observations:	5379	**AIC:**	2.227e+04
Df Residuals:	5344	**BIC:**	2.250e+04
Df Model:	34		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
text_am	0.0013	0.000	10.209	0.000	0.001	0.001
log_n_img	0.4862	0.042	11.652	0.000	0.404	0.568
created	-0.0086	0.033	-0.262	0.793	-0.073	0.056
log_mgp_per_photo	1.4313	0.295	4.849	0.000	0.853	2.010
quest	0.0562	0.077	0.733	0.463	-0.094	0.206
max_pled_t	2.02e-05	1.26e-05	1.604	0.109	-4.48e-06	4.49e-05
is_site	0.4567	0.069	6.581	0.000	0.321	0.593
mmo	-0.3767	0.108	-3.484	0.000	-0.589	-0.165
backed	0.0084	0.003	2.460	0.014	0.002	0.015
strategy	0.0864	0.071	1.216	0.224	-0.053	0.226
simulator	0.2383	0.143	1.665	0.096	-0.042	0.519
period	0.0002	0.003	0.057	0.955	-0.006	0.006
adventure	0.2246	0.060	3.756	0.000	0.107	0.342
n_pled_t	0.0880	0.012	7.211	0.000	0.064	0.112
text_am_sq	-2.317e-07	2.73e-08	-8.481	0.000	-2.85e-07	-1.78e-07
log_min_pled_t	0.1095	0.038	2.917	0.004	0.036	0.183
shooter	-0.1867	0.092	-2.035	0.042	-0.366	-0.007
n_vid_binary	1.6652	0.064	25.818	0.000	1.539	1.792
horror	0.0968	0.090	1.079	0.281	-0.079	0.273
arcade	0.1649	0.090	1.841	0.066	-0.011	0.340
survival	-0.0380	0.080	-0.475	0.635	-0.195	0.119
platformer	0.0371	0.089	0.417	0.676	-0.137	0.211
log_cr_time	0.0711	0.016	4.394	0.000	0.039	0.103
log_step_pled_t	0.0759	0.029	2.635	0.008	0.019	0.132
cont	-0.3987	0.337	-1.182	0.237	-1.060	0.263
indie	0.2080	0.061	3.402	0.001	0.088	0.328
const	-0.3231	0.226	-1.431	0.152	-0.765	0.119
fighting	-0.2546	0.077	-3.295	0.001	-0.406	-0.103
rpg	0.1656	0.065	2.552	0.011	0.038	0.293
log_goal	0.2540	0.023	11.163	0.000	0.209	0.299
action	-0.1306	0.062	-2.120	0.034	-0.252	-0.010
date_year_2013	-0.3303	0.080	-4.130	0.000	-0.487	-0.174
date_year_2014	-0.7115	0.088	-8.096	0.000	-0.884	-0.539
date_year_2015	-0.7791	0.086	-9.010	0.000	-0.949	-0.610
date_year_2016	-0.7171	0.095	-7.555	0.000	-0.903	-0.531

Figure 19: Regression for the equation above (best variables)

So I am going to choose the group of variables and delete/aggregate them with the help of F-test:

1) Aggregating 2014, 2015, 2016 year variables:

F-test: $H_0 : \beta_3 2 = \beta_3 3 = \beta_3 4$

```
<F test: F=array([[0.37861749]]), p=0.6848258786081431, df_denom=5.34e+03, df_num=2>
```

Figure 20: F-test

P-value is 0.685, so the restrictions are valid, null hypothesis is not rejected

2) Deleting insignificant tag variables:

F-test: $H_0 : \beta_2 0 = \beta_2 2 = \beta_1 9 = \beta_1 1 = \beta_5 = \beta_2 1 = \beta_1 0 = 0$

```
<F test: F=array([[1.20072146]]), p=0.2984525824160479, df_denom=5.35e+03, df_num=7>
```

Figure 21: F-test

P-value is 0.298, so the restrictions are valid, null hypothesis is not rejected

3) Deleting other insignificant variables:

F-test: $H_0 : \beta_3 = \beta_6 = \beta_1 2 = \beta_2 5 = 0$

```
<F test: F=array([[1.03883721]]), p=0.3854985131891523, df_denom=5.35e+03, df_num=4>
```

Figure 22: F-test

P-value is 0.385, so the restrictions are valid, null hypothesis is not rejected

So final preliminary model is as follows: $log\_pled_i = \beta_0 + \beta_1 text\_am_i + \beta_2 log\_n\_img_i + \beta_3 log\_mgp\_per\_photo_i + \beta_4 is\_site_i + \beta_5 mmo_i + \beta_6 backed_i + \beta_7 adventure_i + \beta_8 n\_pled\_t_i + \beta_9 text\_am\_sq_i + \beta_{10} log\_min\_pled\_t_i + \beta_{11} shooter_i + \beta_{12} n\_vid\_binary_i + \beta_{13} log\_cr\_time_i + \beta_{14} log\_step\_pled\_t_i + \beta_{15} indie_i + \beta_{16} fighting_i + \beta_{17} rpg_i + \beta_{18} log\_goal_i + \beta_{19} action_i + \beta_{20} date\_year\_2013_i + \beta_{21} date\_year\_2014\_15\_16_i + v_i$

And the estimation of it:

Dep. Variable:	log_pled	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	342.6
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.00
Time:	12:46:09	Log-Likelihood:	-11108.
No. Observations:	5379	AIC:	2.226e+04
Df Residuals:	5357	BIC:	2.240e+04
Df Model:	21		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
date_year_2014_15_16	-0.7172	0.072	-9.930	0.000	-0.859	-0.576
log_min_pled_t	0.1164	0.037	3.138	0.002	0.044	0.189
shooter	-0.1867	0.091	-2.062	0.039	-0.364	-0.009
n_vid_binary	1.6705	0.064	25.986	0.000	1.545	1.797
text_am	0.0013	0.000	10.487	0.000	0.001	0.002
log_cr_time	0.0691	0.016	4.365	0.000	0.038	0.100
log_n_img	0.4581	0.036	12.748	0.000	0.388	0.529
log_mgp_per_photo	1.2928	0.253	5.106	0.000	0.797	1.789
log_step_pled_t	0.1045	0.022	4.687	0.000	0.061	0.148
indie	0.2158	0.061	3.560	0.000	0.097	0.335
is_site	0.4603	0.069	6.694	0.000	0.326	0.595
mmo	-0.3730	0.107	-3.480	0.001	-0.583	-0.163
log_goal	0.2633	0.023	11.692	0.000	0.219	0.307
const	-0.5256	0.185	-2.844	0.004	-0.888	-0.163
fighting	-0.2490	0.077	-3.231	0.001	-0.400	-0.098
backed	0.0086	0.003	2.595	0.009	0.002	0.015
date_year_2013	-0.3267	0.080	-4.092	0.000	-0.483	-0.170
adventure	0.2183	0.059	3.722	0.000	0.103	0.333
n_pled_t	0.0916	0.012	7.844	0.000	0.069	0.114
rpg	0.1576	0.064	2.471	0.013	0.033	0.283
text_am_sq	-2.292e-07	2.69e-08	-8.525	0.000	-2.82e-07	-1.77e-07
action	-0.1168	0.061	-1.912	0.056	-0.237	0.003

Figure 23: Estimation of preliminary final model