

Code, Quality, and Process Metrics in Graduated and Retired ASFI Projects

Ștefan Stănciulescu

sstanciulescu@ucdavis.edu

University of California, Davis
USA

Likang Yin

lkyin@ucdavis.edu

University of California, Davis
USA

Vladimir Filkov

vfilkov@ucdavis.edu

University of California, Davis
USA

In Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp 495–506

OSSプロジェクトの持続可能性

- | 昨今のデジタル社会において、プロジェクトの持続可能性は重要なテーマ
- | 多くの新興プロジェクトにとって、持続可能性の確保や長期プロジェクト維持が課題となる

- 持続可能なOSSプロジェクトの特徴は何か？
- プロジェクト維持にはどのように注力できるのか？

目的

プロジェクトの持続可能性に関係する特徴を定量的に明らかにする

- ソースコード, コーディングプロセス, プロジェクト品質

ASFIに属していたプロジェクトを対象に, 卒業したプロジェクトと除籍されたプロジェクトの違いを分析

- ASFI (Apache Software Foundation Incubator)
 - 持続可能なプロジェクトを作るための支援を行う団体
- 卒業
 - 持続可能なプロジェクトとして認定されること
- 除籍
 - 持続可能性が乏しいために支援の対象外となること

Research Question

RQ1

- ASFIから卒業または除籍されたプロジェクトでメトリクスはどのように異なっているか？
 - 統計値での比較

RQ2

- ASFIから卒業または除籍されたプロジェクトの軌跡はメトリクス沿って異なるか？
 - 時系列での比較

RQ3

- ASFIから卒業または除籍するかを決定づけるメトリクスにはどのようなものがあるか？
 - 回帰モデルのオッズ比などによる比較

データセット

GitHubに存在する236プロジェクト

- ASFIによる支援を受けたことがある, もしくは, 現在受けているプロジェクトをpodings.xmlから取得
 - <https://svn.apache.org/repos/asf/incubator/public/trunk/content/podlings.xml>
- SVNリポジトリを所有するプロジェクトはリポジトリの特定ができないため除外

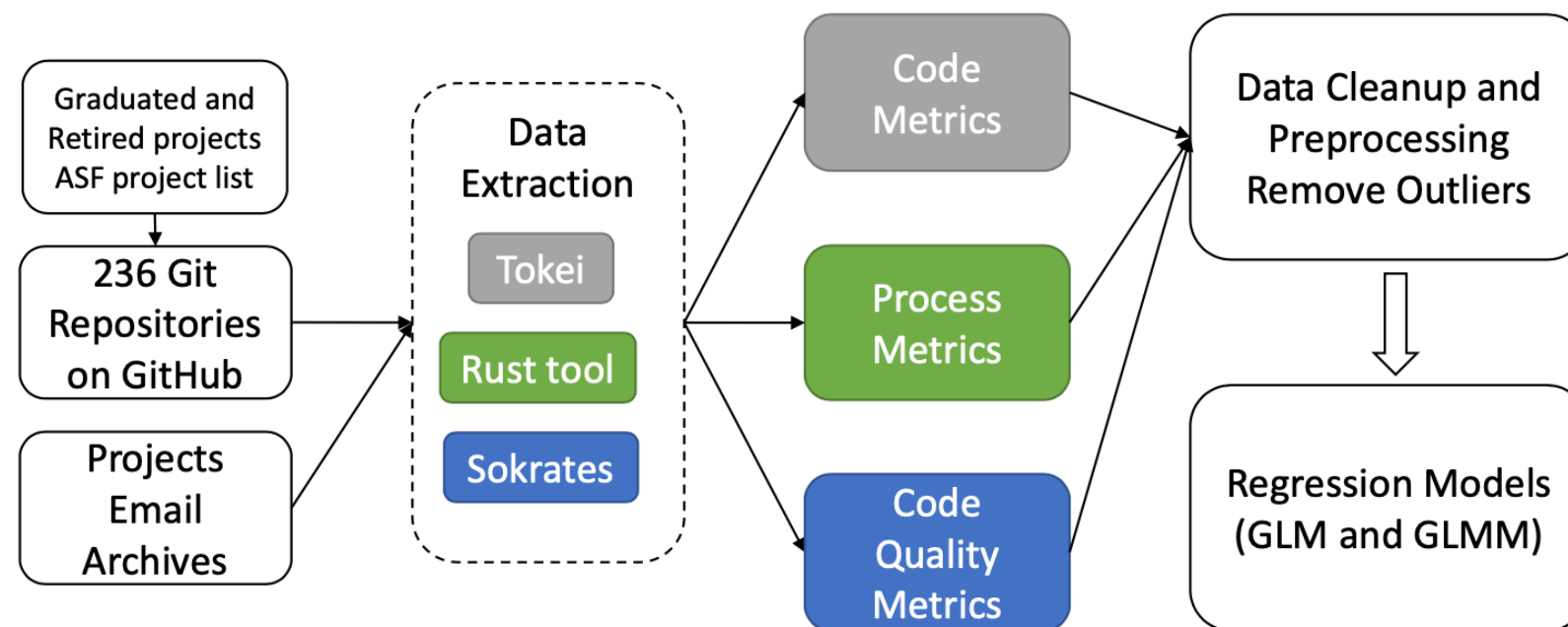
利用したプロジェクト数

種類	プロジェクト数
ASFIを卒業	184
ASFIを除籍	34

手順

リポジトリデータとメーリングアーカイブを利用し、下記のツールにてメトリクスを計測

- [tokei](#)¹, [Rust tool](#)², [Sokrates](#)³



計測したメトリクスを用いて回帰モデルの構築

- 一般化線形回帰モデルと一般化線形混合効果モデル

1. github.com/XAMPPRocky/tokei
2. <https://zenodo.org/record/6374071>
3. <https://www.sokrates.dev/>

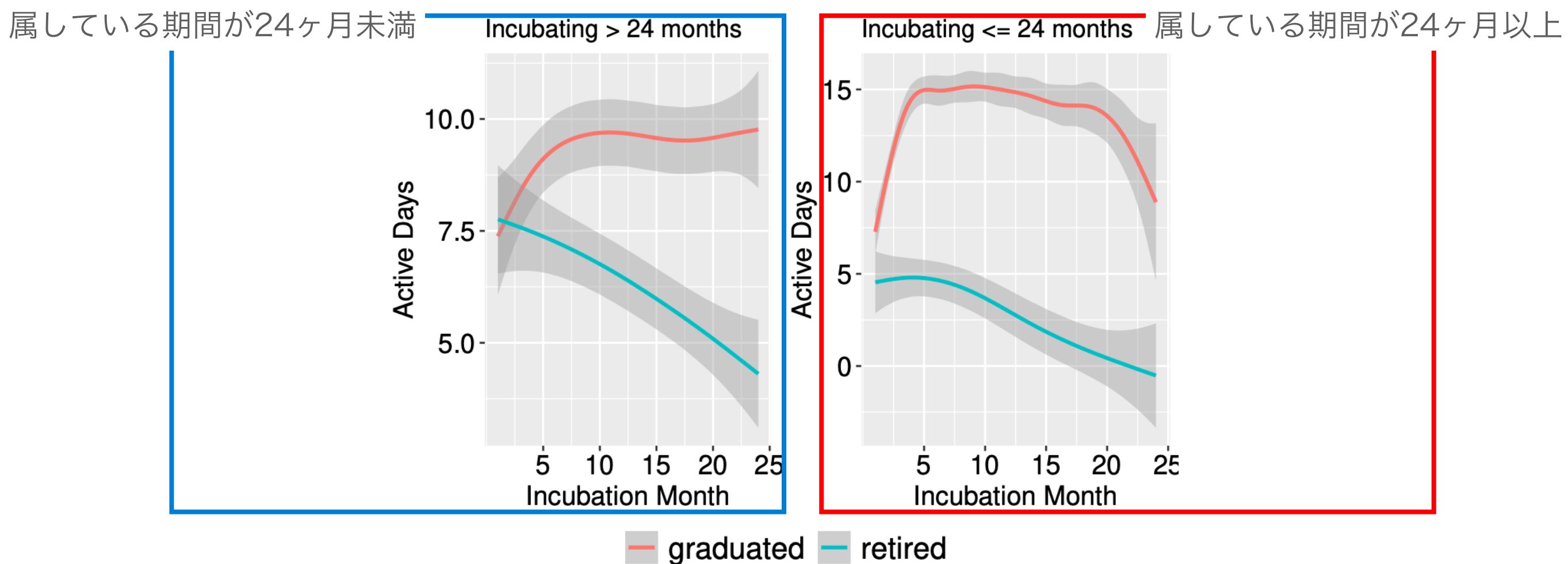
RQ1：メトリクスごとの違い

- ほとんどのメトリクスで違いが見られた

	184 graduated projects			34 retired projects		
	mean	median	st.dev	mean	median	st.dev
SLOC	123431.66	66219.00	200938.97	89866.66	37513.00	145736.58
Commits	50.36	22.00	90.96	11.39	2.00	23.70
Directories	435.19	222.00	583.98	311.33	176.00	316.17
Top Level Directories	10.31	8.00	10.26	9.53	8.00	7.22
Incubation Months	15.98	12.00	13.42	24.71	21.00	17.82
Authors	5.69	3.00	7.88	1.52	1.00	1.96
Major Contributors	2.92	3.00	2.18	1.29	1.00	1.36
Minor Contributors	2.73	0.00	6.73	0.21	0.00	0.85
New Contributors	1.65	0.00	3.72	0.37	0.00	1.03
Files Added	172.45	18.00	807.46	56.27	0.00	432.19
Files Deleted	131.30	3.00	735.82	37.94	0.00	379.24
Files	957.01	609.00	1077.36	807.78	371.00	1227.41
Avg Files Modified per Commit	5.72	3.33	11.78	5.26	1.49	23.14
Active Days	10.64	9.00	8.55	3.83	2.00	5.38
Emails	131.88	63.00	217.79	42.48	16.00	79.04
Corrective	11.60	5.00	18.56	2.47	0.00	5.65
Features	14.07	6.00	25.52	3.27	0.00	7.04
Perfective	23.34	9.00	53.25	5.41	1.00	11.98
Non Functional	0.86	0.00	1.83	0.17	0.00	0.59
Number of functions	3570.97	2350.00	3934.19	2768.23	1384.50	3888.07
Test/Main Lines of Code Percentage	47.43	27.56	126.30	32.87	23.41	36.01
#Functions /w McCabe Index 11-25	81.81	47.00	107.22	52.74	24.00	63.94
#Functions /w McCabe Index >51	2.47	0.00	5.41	0.82	0.00	1.69
Very Large File Size Count	6.89	1.00	21.97	3.40	1.00	4.69
Very Large Function Size Count	17.17	5.00	32.10	9.74	3.00	15.41
Code Duplication Percentage	16.38	12.69	14.92	16.93	13.63	15.69
Most Complex Function LOC	305.67	134.00	1193.30	178.81	120.00	156.75

RQ2：プロジェクトごとの軌跡の違い

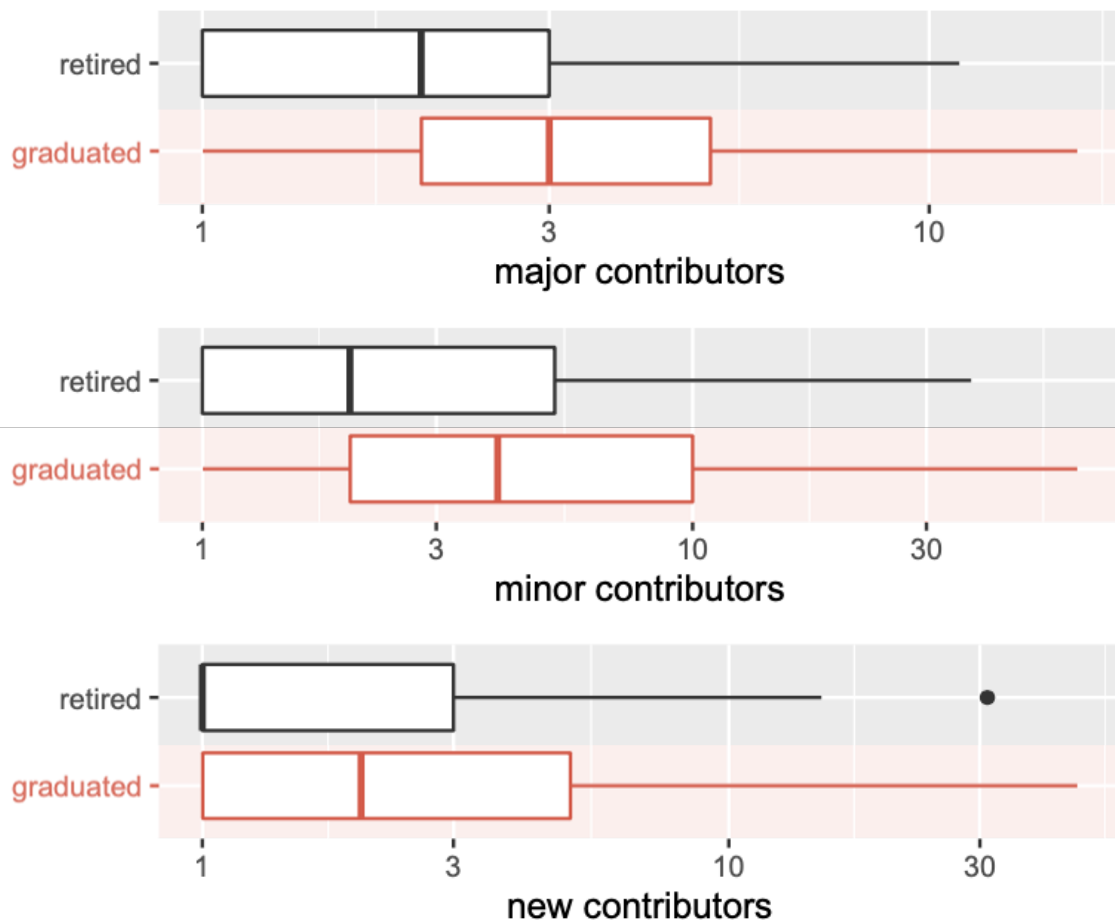
- ASFIに属している期間が24ヶ月未満と以上での違い
(平均24ヶ月でASFIを除籍される)



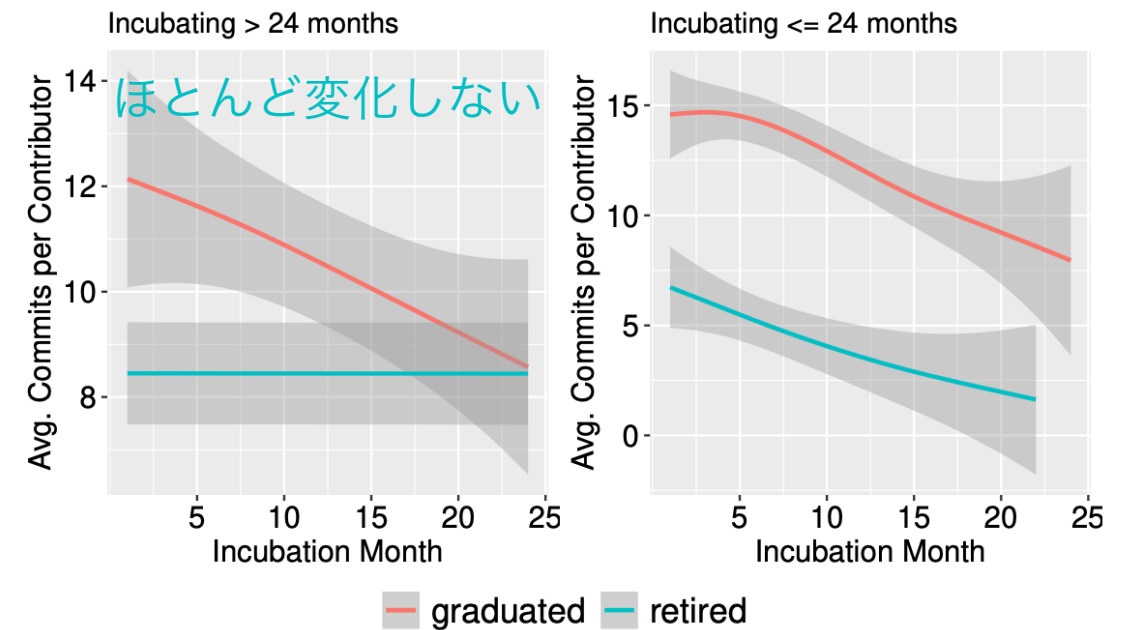
- ASFIを卒業するプロジェクトはアクティブな日数が多い
- 除籍されるプロジェクトはアクティブな日数が減少する傾向

RQ2：プロジェクトごとの軌跡の違い

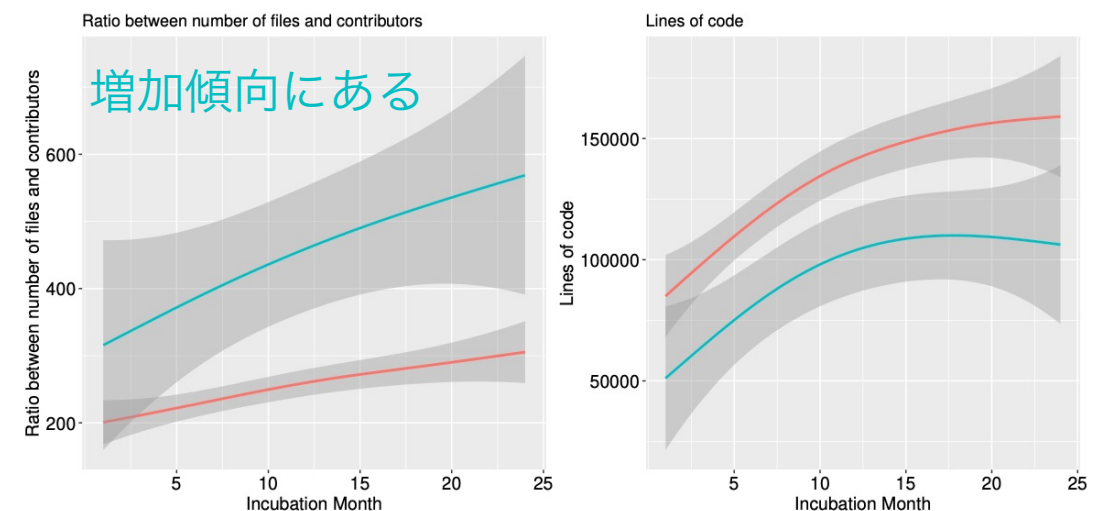
開発者の数



開発者一人当たりのコミット数



開発者とファイル数の比率



- ASFIを卒業するプロジェクトは**開発者の数が多い**
- 除籍されるプロジェクトは開発者の数が少ないため、一人当たりの負担が大きい

RQ3：プロジェクトごとの軌跡の違い

ポジティブな影響を与える因子

ネガティブな影響を与える因子

Predictors	Base		Base, Processes		Base, Processes, Quality	
	Odds Ratios	p	Odds Ratios	p	Odds Ratios	p
Intercept	0.69	0.491	21.31	<0.001	60.25	<0.001
SLOC	1.21	<0.001	0.93	0.129	0.98	0.759
#Directories	1.27	<0.001	1.22	0.002	1.15	0.045
#Top Level Directories	0.67	<0.001	0.63	<0.001	0.65	<0.001
Incubation Month	0.53	<0.001	0.73	<0.001	0.70	<0.001
#Major Contributors			2.15	<0.001	2.21	<0.001
#Minor Contributors			1.09	<0.001	1.09	<0.001
#New Contributors			0.98	0.269	0.98	0.264
#Files Added			1.00	0.764	1.01	0.675
#Files Deleted			0.96	0.001	0.96	0.003
Avg. Files Modified per Commit			0.93	0.001	0.92	<0.001
Active Days			1.35	0.015	1.30	0.036
#Emails			0.97	0.078	0.98	0.179
#Corrective			1.06	0.001	1.05	0.005
#Features			1.06	<0.001	1.06	<0.001
#Perfective			1.00	0.969	1.00	0.903
#Non Functional			1.03	0.138	1.03	0.159
Test/Main Lines of Code Percentage					1.04	0.079
#Functions /w McCabe Index 11-25					1.01	0.772
#Functions /w McCabe Index >51					1.01	0.446
Very Large File Size Count					0.90	<0.001
Very Large Function Size Count					1.19	<0.001
Code Duplication Percentage					0.87	0.004
Most Complex Function LOC					0.73	<0.001
Observations	4970		4970		4970	
Marginal R2 / Conditional R2	0.097 / 0.539		0.213 / 0.689		0.154 / 0.844	

RQ3

ポジティブな影響を持つ因子

- 開発者数, アクティブな日数, バグ修正コミット数や機能追加コミット数
- 開発者数やバグ修正コミットの増加はプロジェクトの保守性を向上させ持続可能性を高める

ネガティブな影響を持つ因子

- 1コミットあたりの変更ファイル数, コードの重複割合, サイズ (LOC) が大きいファイル数
- コードベース (やソースファイル) が複雑であるほど保守しにくくなり持続可能性が低下する

まとめ

目的

- OSSプロジェクトの持続可能性に影響する要因を定量的に明らかにする

貢献

- コード，プロセス，品質の観点から，プロジェクトの持続可能性に関係する要因を明らかにした
- ASFIを卒業したプロジェクトと除籍されたプロジェクトの軌跡の違いを明らかにした
- ASFIを卒業するプロジェクトの予測に影響する要因を明らかにした

Javaプロジェクトでの結果

	Base		Base, Processes		Base, Processes, Quality	
Predictors	Odds Ratios	p	Odds Ratios	p	Odds Ratios	p
Intercept	0.83	0.378	3.49	< 0.001	1.85	0.120
SLOC	1.22	< 0.001	1.11	0.001	1.15	0.004
#Top Level Directories	0.79	< 0.001	0.80	< 0.001	0.79	0.001
Incubation Month	0.48	< 0.001	0.63	< 0.001	0.59	< 0.001
#Major Contributors			1.82	< 0.001	1.84	< 0.001
#Minor Contributors			1.06	0.012	1.05	0.022
#New Contributors			0.97	0.151	0.98	0.227
#Files Added			1.02	0.372	1.02	0.245
#Files Deleted			0.96	0.010	0.97	0.047
Avg. Files Modified per Commit			0.95	0.016	0.94	0.004
#Emails			0.96	0.011	0.96	0.052
#Corrective			1.08	< 0.001	1.07	0.001
#Features			1.08	< 0.001	1.08	< 0.001
#Perfective			1.00	0.963	0.98	0.453
#Non Functional			0.99	0.557	0.99	0.757
Test/Main Lines of Code Percentage					1.01	0.713
#Functions /w McCabe Index 11-25					1.10	0.002
#Functions /w McCabe Index >51					1.01	0.408
Very Large File Size Count					0.81	< 0.001
Very Large Function Size Count					1.18	< 0.001
Code Duplication Percentage					0.85	< 0.001
Observations	3795		3795		3795	
R2 Tjur	0.116		0.185		0.239	

一般化線形回帰モデルと一般化線形混合効果モデル

- 一般化モデルはバイナリ変数による影響を低減させる仕組みが含まれている → ロジスティック回帰
- 一般化線形混合効果モデルは、観測データの個体差による影響を考慮するためのモデル
 - 通常、固定効果のみでの評価となるが、一般化線形混合効果モデルは変量効果と呼ばれる項が追加され、非正規分布による影響などを軽減する
 - 今回はプログラミング言語を変量効果として設定し、開発する製品の違いによる影響を低減している