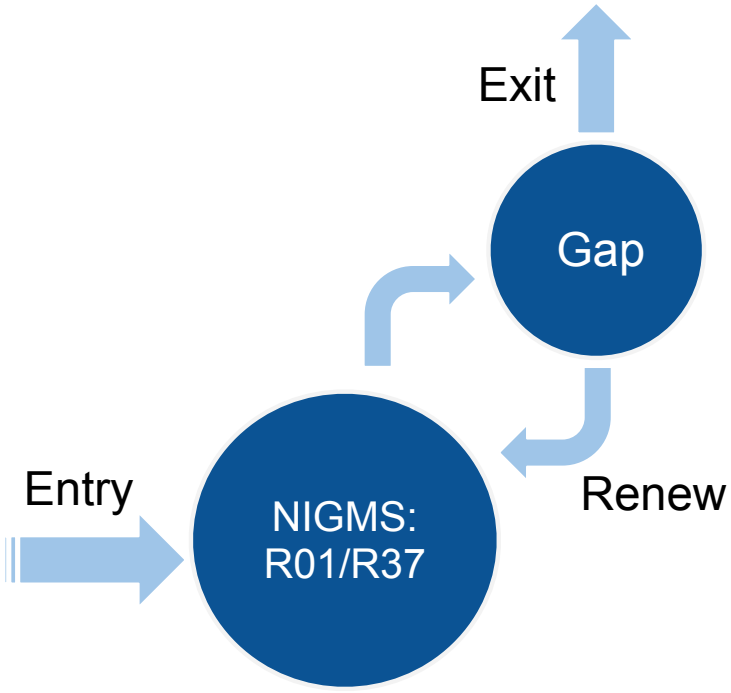
Understanding the Factors that Modulate the Biomedical Research Workforce

NIH Mentors:
Jake Basson, Lisa Hechtman, Anna Calcagno
Team Members:
Jingning Li, Xin Yuan
Xinyi Wang, Yunning Zhu

Introduction



Graph: NIGMS R01/R37 grantees' flow.

Organization

NIGMS -The National Institution of General Medical Sciences

Background and Goals

- Getting funded is never easy for members of the biomedical research workforce
- NIH has found that when researchers experience time without funding, the longer researchers stay unfunded, the less likely they are to return to the NIH workforce pool
- By digging deeper into NIH grants records, we aim to find key factors to better explain and predict when investigators lose or gain funding
- The findings will help NIGMS efficiently allocate biomedical funds for long-term planning purposes as well as identify better ways to preserve and support researchers

Problem Identification

Important Concept - *Gap*
A situation where a project is temporarily or permanently not funded.

Reason for Studying Gaps
- When investigators experience funding gaps, they encounter financial challenges.
- If an investigator is not able to continue research, NIH also loses their investment.

Other Potential Factors
- Stability of funding
- The number of concurrent projects/funding
- Number of support years

...

Project Workflow, Methods & Tools

Workflow:

Methods:

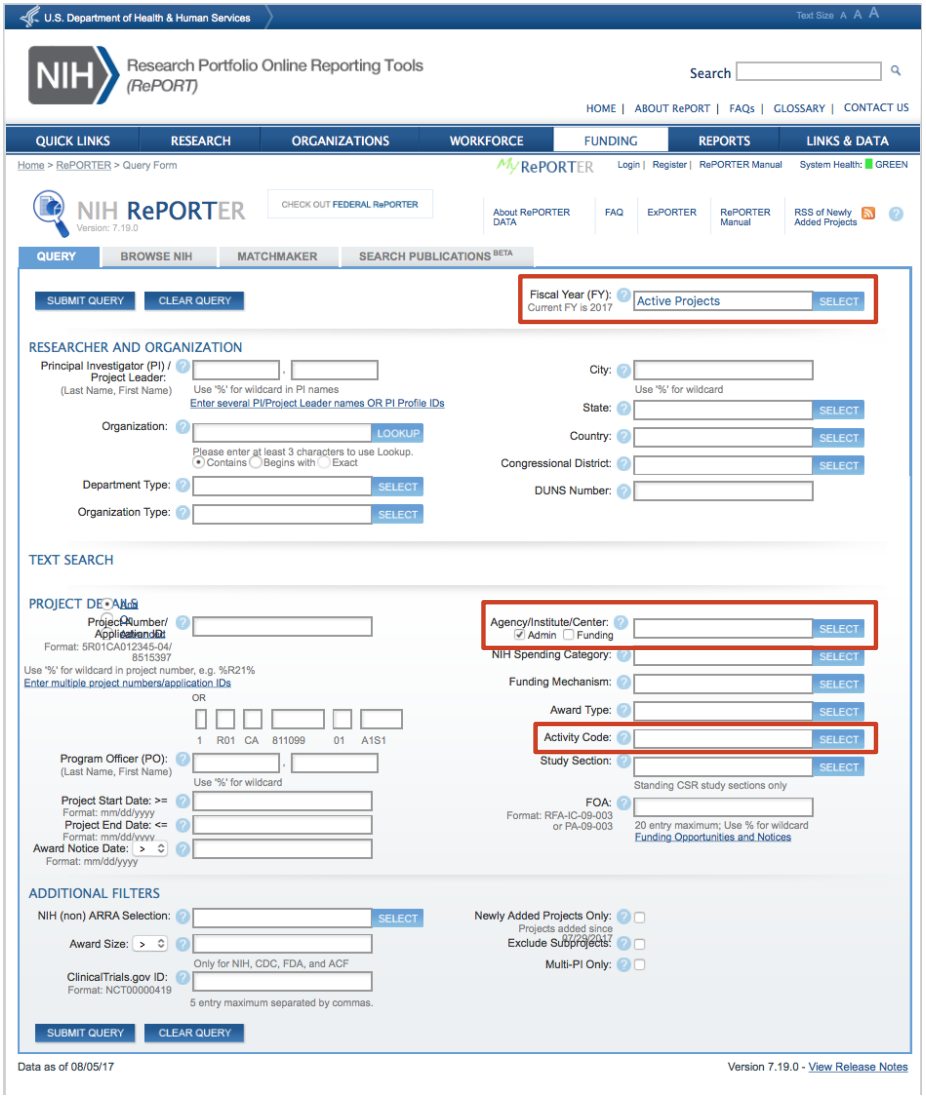
Logistic Regression (Primary), Multilevel Logistic Regression & Stepwise Model Selection (Complementary)

Tools:

R, Python, JMP, Fuzzy Match/Excel

Data Overview

- Main Data Source: [Public Grants Database (NIH/ Federal RePORTER)](), containing annual records of funded projects
- Time Span: From 1993 to 2015



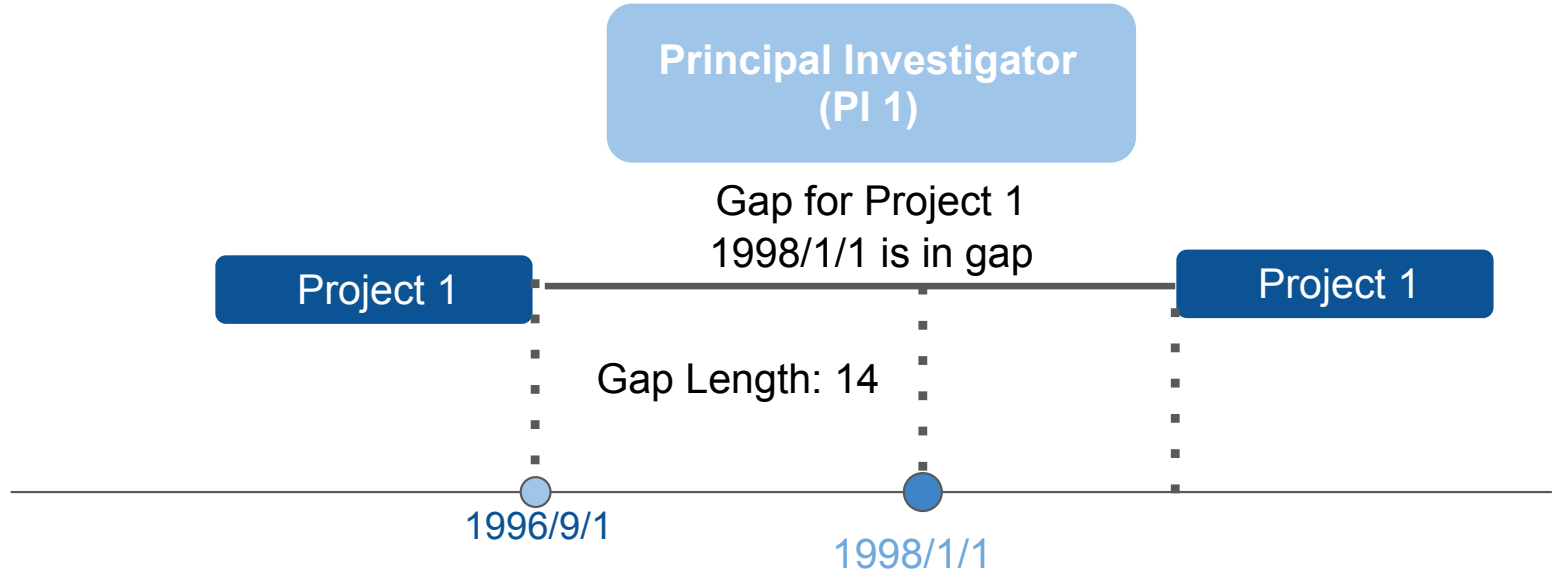
Data Overview

	PPID	pjnum	Gap.Length	SY	FY	idea	Carn15	Status6	Status12	Status18	Status24	Status30	Status36	StatusAny	SY2	Num.Concurrent	Funding.Concurrent
1	1857698	39586	40.07671233	6	1993	0	NA	0	0	0	0	0	0	0	36	0	0.000
2	1857700	36344	9.07397260	9	1994	0	1	0	0	0	0	0	0	0	81	0	0.000
3	1857700	47922	0.03287671	1	1994	0	1	0	0	0	0	0	0	0	1	0	0.000
4	1857701	33324	6.08219178	4	1993	0	0	0	0	0	0	0	0	0	16	0	0.000
5	1858038	48807	40.07671233	1	1994	0	1	1	1	1	1	1	1	1	1	1	247.149
6	1858040	48449	8.08767123	1	1994	0	1	0	0	0	0	0	0	0	1	0	0.000
7	1858043	46812	17.06301370	1	1993	1	NA	0	0	0	0	0	0	0	1	1	154.403
8	1858045	47453	0.03287671	1	1995	0	0	0	0	0	0	0	0	0	1	0	0.000
9	1858048	42680	5.06301370	5	1993	0	1	0	1	1	1	1	1	1	25	0	0.000
10	1858048	46749	17.06301370	5	1996	0	1	0	0	0	0	0	0	1	25	0	0.000
11	1858052	38133	6.08219178	6	1993	0	NA	1	1	1	1	1	1	1	36	0	0.000
12	1858054	32707	6.08219178	9	1993	0	NA	1	1	1	1	1	1	1	81	0	0.000
13	1858055	30147	6.08219178	12	1993	0	1	0	0	0	0	0	0	0	144	0	0.000
14	1858057	44669	10.06027397	4	1994	0	NA	0	0	0	0	0	0	0	16	1	136.749
15	1858374	48129	21.07397260	1	1994	0	NA	0	0	0	0	0	0	0	1	1	262.194
16	1858378	41679	9.07397260	5	1993	0	1	0	0	0	0	0	0	0	25	1	171.802
17	1858388	55372	4.04383562	1	1997	0	NA	0	0	0	0	0	0	1	1	0	0.000
18	1858392	47299	16.04383562	1	1993	0	NA	0	0	0	0	0	0	0	1	0	0.000
19	1858395	49988	20.08767123	1	1994	0	NA	0	1	1	1	1	1	1	1	1	200.681

Showing 1 to 20 of 3,847 entries

Key Concepts

Let's take an example

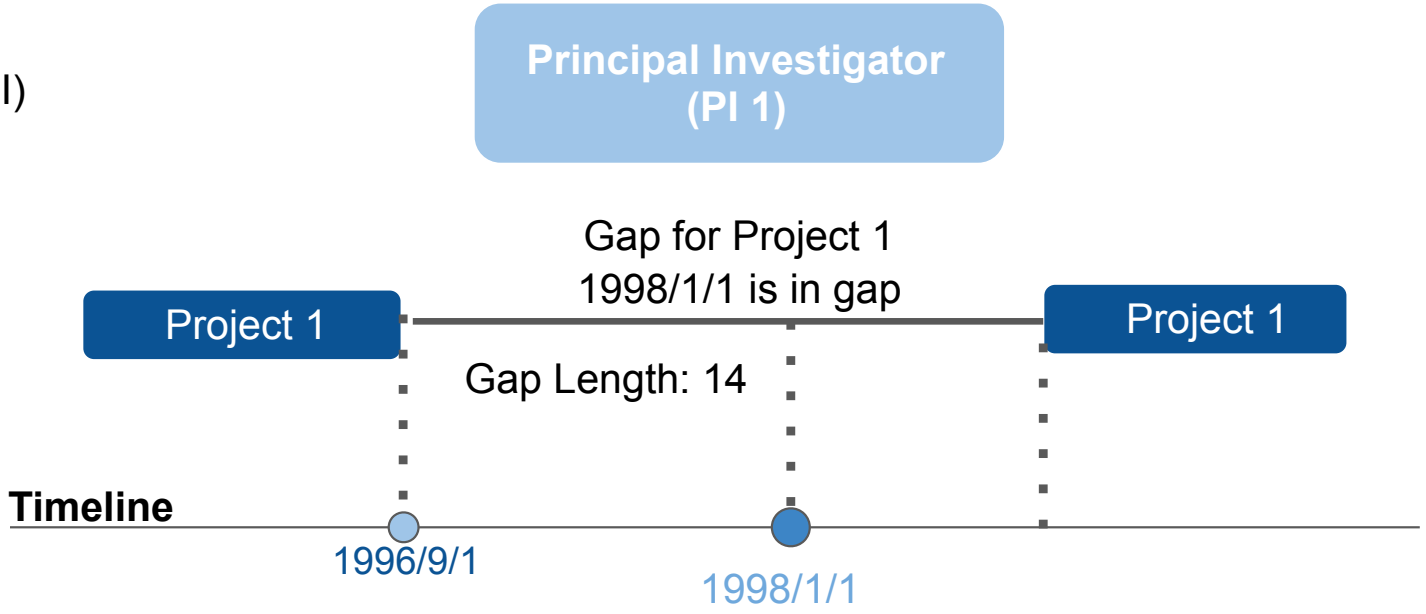


Principal Investigator (PI)	Project (PJ)	Gap Length
PI 1	PJ 1	14

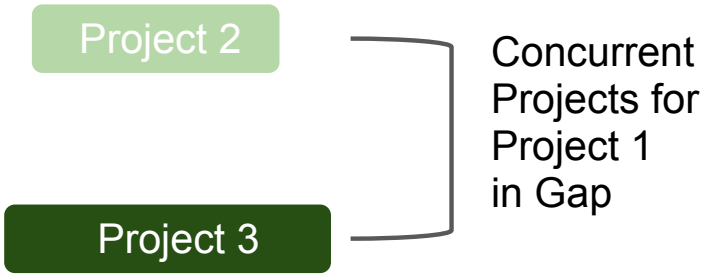
- Most gaps last no longer than 4 years
- Judge if one project is in gap on four specific dates:
  Model Construction: 1998/1/1; 2002/1/1; 2006/1/1
  Model Validation: 2010/1/1

Key Concepts

- One principal investigator (PI) can have several ongoing projects at the same time
- Funding for Project 2: 100k
  Funding for Project 3: 200k
- PIs with more concurrent projects tend to have more motivation to apply for more fundings



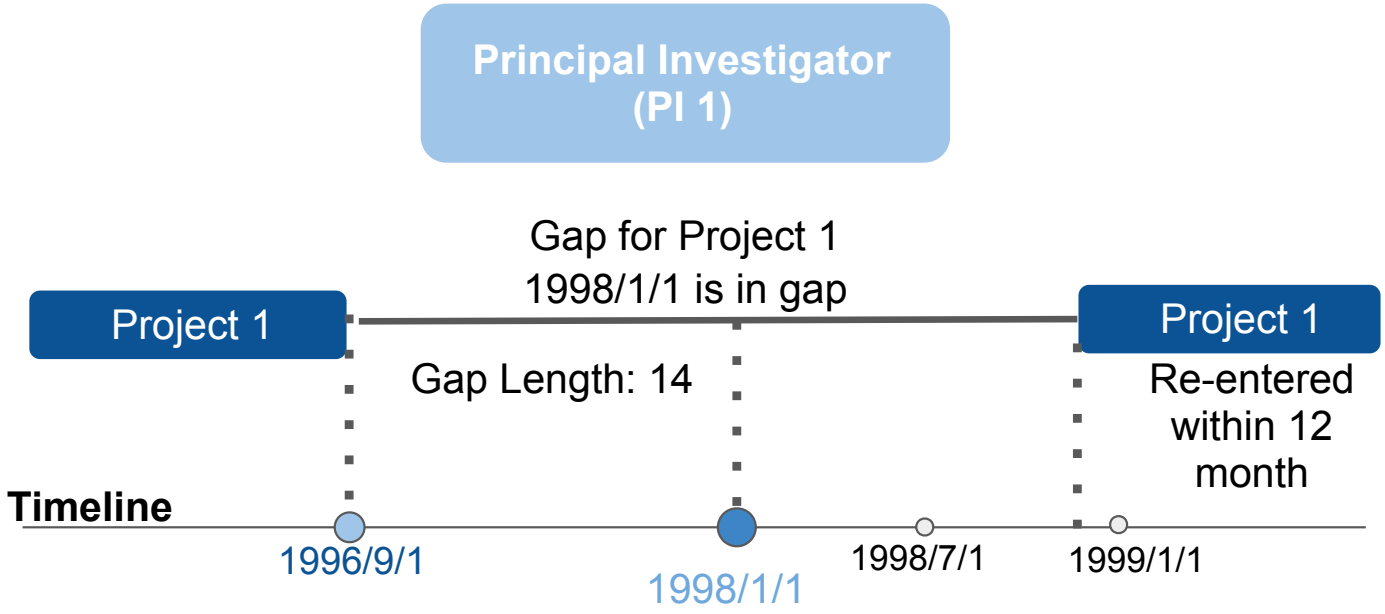
Principal Investigator (PI)	Project (PJ)	Gap Length	Concurrent Projects	Concurrent Funding
PI 1	PJ 1	14	2	100k + 200k



Key Concepts

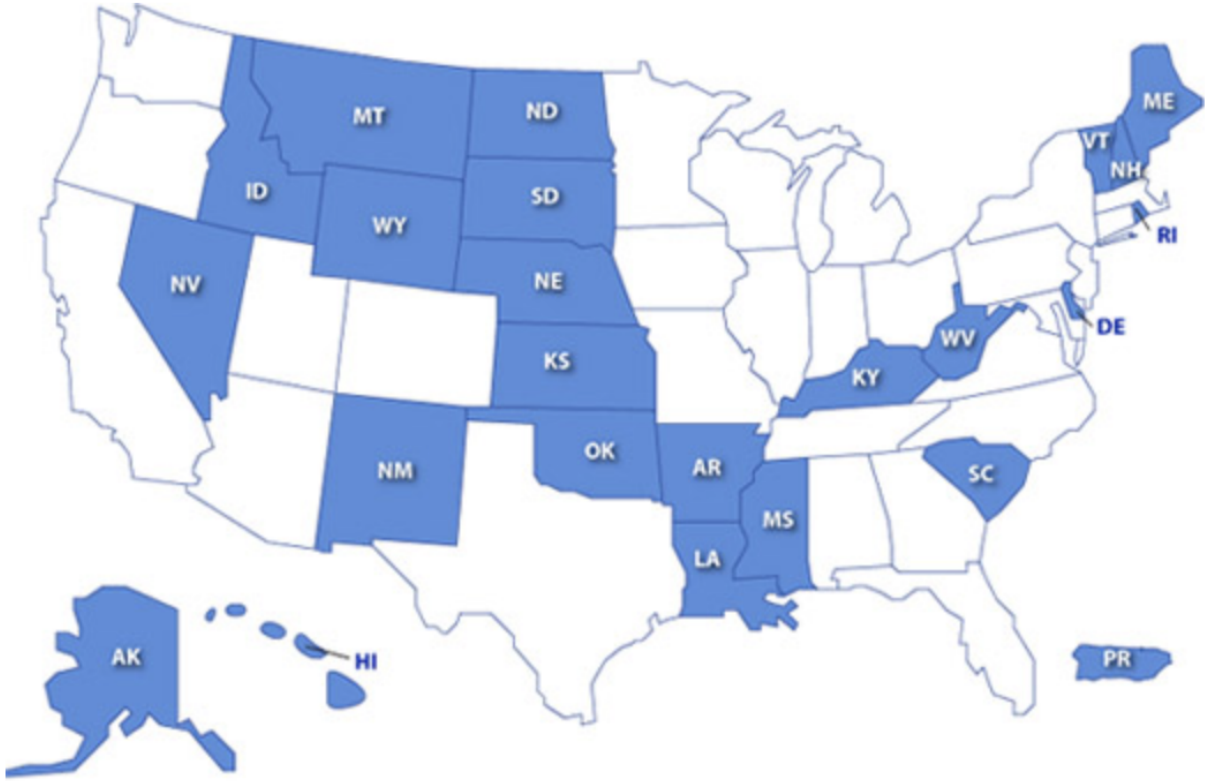
Targets

- Status 6/12/.../36:
  - 1 ---- Re-entered within next 6/12/.../36 months
  - 0 ---- Hasn't re-entered within next 6/12/.../36 months
- Status Any:
  - 1 ---- Re-entered after gap
  - 0 ---- Still in gap



Principal Investigator (PI)	Project (PJ)	Gap Length	Concurrent Projects	Concurrent Funding	Status 6	Status 12	...	Status 36	Status Any
PI 1	PJ 1	14	2	100k + 200k	0	1		1	1

Potential Predictors



Shaded states are eligible for IDeA.

Institutional Development Award (IDeA)

- *"To ...enhance the competitiveness of investigators at institutions located in states in which the aggregate success rate for applications to NIH has historically been low"*
- IDeA: 1/0 (IDeA State/ not IDeA State)

Potential Predictors

Carnegie Classification of Institutions of Higher Education

- A classification system describing research activity at an institution:
- Very high research activity; High research activity; Medical focus

BestName	BASIC2010
Art Academy of Cincinnati	30
Cincinnati College of Mortuary Science	32
Ohio Christian University	22
CUYAHOGA COMMUNITY COLLEGE	7
Academy of Court Reporting and Technology-Cleveland	10
Cleveland Institute of Electronics	10
Ohio Technical College	10
Remington College-Cleveland Campus	10
CASE WESTERN RESERVE UNIVERSITY	15
CLEVELAND CLINIC LERNER COM-CWRU	15

Support Year:

The funding year number that the project is on:

- New project: support year 1
- Ongoing project: support year 2, 3, etc.

Fiscal Year:

Government funding year

Model Construction *- logistic regression*

Target

- **Re-entry** - Binary variable. Value 1 means the project re-enter the funding pool and 0 means otherwise.

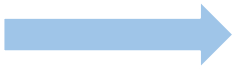
Predictors

- **Support Year** - Years the project has lasted.
- **Gap Length** - Months the project has been unfunded.
- **Concurrent Funding** - The amount of total funding an investigator has on other projects while the current project is in a gap.
- **Concurrent Project** - The number of total concurrent projects a project has during its gap.
- **IF is IDeA State** - Binary variable, value 1 means the research institute is located in an IDeA state, 0 means otherwise.
- **IF Carnegie 15** - Binary variable, value 1 means institute Carnegie code equals to 15, 0 means otherwise.
- **Fiscal Year** - A numeric value represents the fiscal year of project.

Model Construction

Preliminary Predictors
Support Year
Gap Length
Concurrent Funding
of Concurrent Projects
If is IDeA State
If Carnegie 15
Fiscal Year

Improvement



- *Num.Concurrent* and *Funding.Concurrent* have strong positive correlation: 0.92
- *SY* has a quadratic, rather than linear, effect on Gap.Status

Final Predictors
Support Year
(Support Year)^2
Gap Length
Fiscal Year
of Concurrent Projects

Final Model

Probability of Re-Entry ~ Support Year + Support Year^2 + Gap Length + Number of Concurrent Projects + Fiscal Year

Coefficients of Standardized Predictors

	6 Months	12 Months	24 Months	36 Months	Any Months
Concurrent Projects	0.11	0.12	0.06	0.05	0.08
Support Year	0.61	0.66	0.72	0.8	0.98
(Support Year)^2	-0.5	-0.55	-0.6	-0.66	-0.71
Gap Length (months)	-1.12	-1.21	-1.09	-1.01	-0.96
Fiscal Year	-0.15	-0.1	-0.08	-0.07	-0.01

Model Interpretation

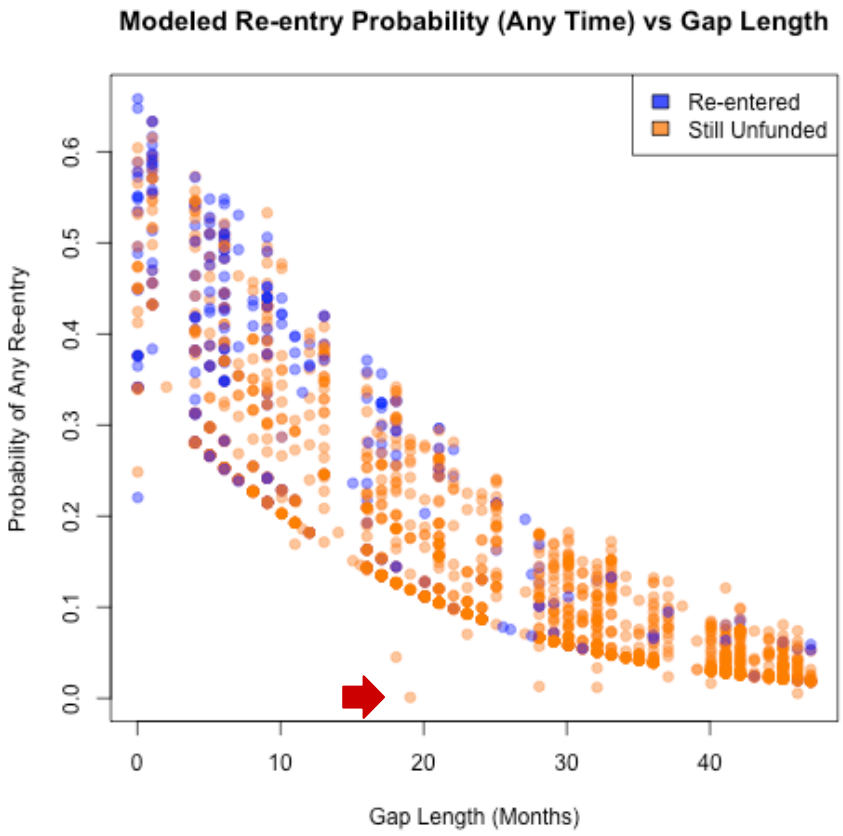
In order of the importance to probability of re-entry, controlling for other variables:

- One-unit increase in **Gap Length** will lead to **8.556% of decrease** in the relative probability of re-entry within the next 12 months;
  ***Reason**: Longer gap means weaker projects that are more difficult to be funded again.*

- One-unit increase in **Fiscal Year** will lead to **3.422% of decrease** in the relative probability of re-entry within the next 12 months;
  ***Reason**: NIH are more strict to applicants in recent years than before.*

- One-unit increase in **Number of Concurrent Project** will lead to **23.759% of increase** in the relative probability of re-entry within the next 12 months;
  ***Reason:** Investigators with other work going on have a financial buffer which could help them keep running their lab while applying to renew the project that's in a gap.*

Model Interpretation



A closer look at the effect of Gap Length:

- **Blue points**: projects that really returned after gap
- **Yellow points**: projects that really didn't come back

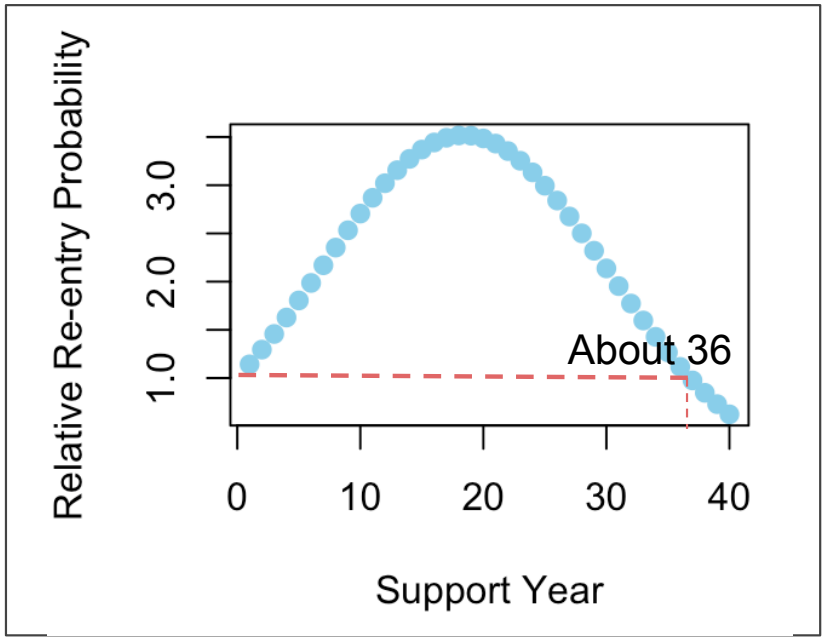
X-lab: Gap length in number of months
Y-lab: Re-entry probability model predicted

we could conclude:

· For those projects that did re-enter after gap, our model gave a higher estimate of re-entry probability
· Majority of returned projects have a gap length no longer than 15 months

Model Interpretation

· **Support Year** : SY , SY^2



Graph: effect of SY on Relative re-entry probability

Support Year has a quadratic effect on the Odds of *Gap.status.*

Turning point - about **20** years :

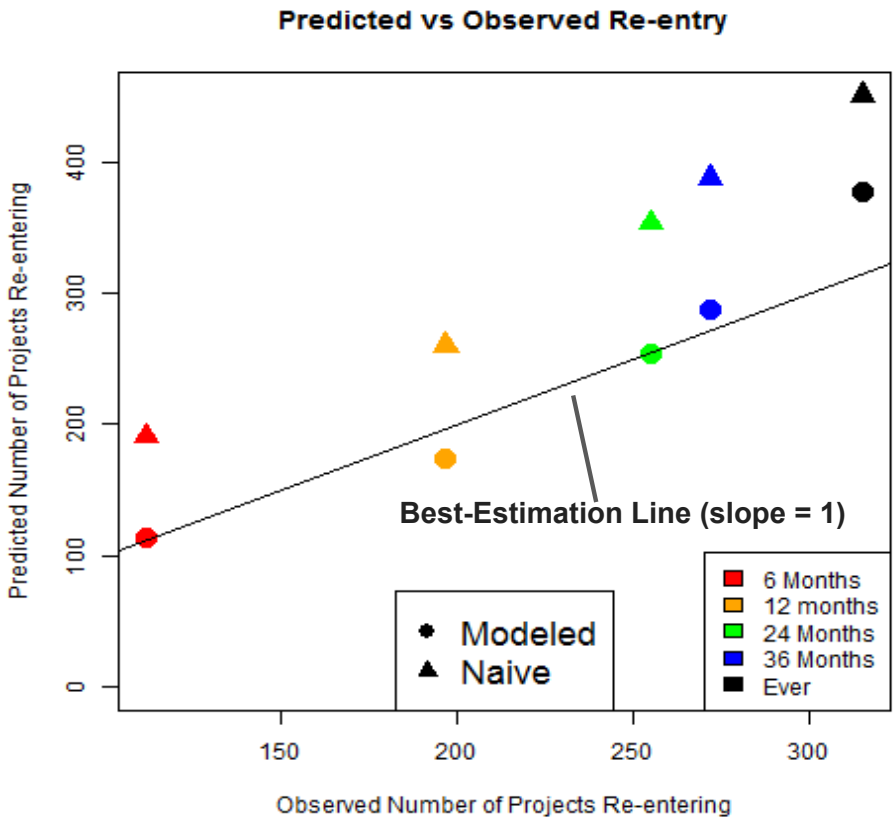
· Before: positive effect

Reasons: - *Unsuccessful projects will end earlier, while more successful projects (longer ones) are more likely to be funded again*

· After: negative effect

Reasons: - *Researcher's age*
- *New research topics identified*

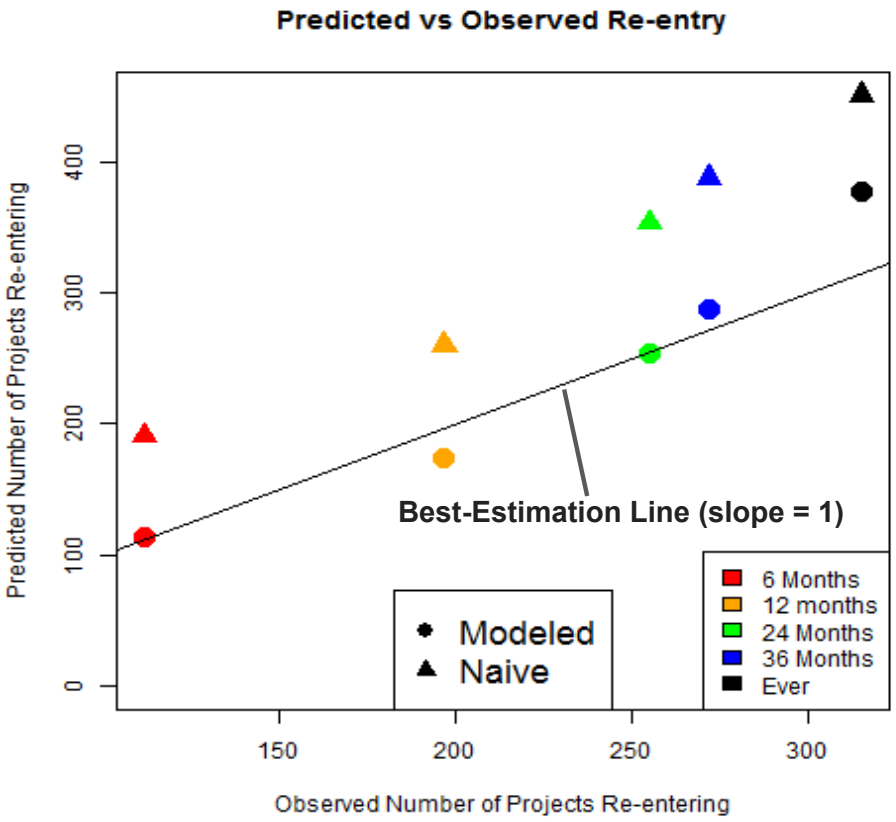
Model Validation - 2010 Dataset



▲ **Value of Naive Points** = Real Returning Fraction in 1998-2006 * Number of Records in 2010 Dataset

● **Value of Modeled Points** = Predicted Returning Fraction in 2010 * Number of Records in 2010 Dataset

Model Validation - 2010 Dataset



Large Distances between Modeled Points and Naive Points

Modeled-points are scattering around the Best-Estimation Line

Summary & Recommendations

Gap Length, **Support Year**, **Number of Concurrent Projects** and **Fiscal Year** are key factors, which significantly predict the likelihood that a project will be funded again.

- The Larger the Gap Length is, the harder it would be for NIH to sustain projects that have had gaps
- The relationship between Support Year and re-entry probability depends on the number of Support Years. Up until about support year 20, the relationship is positive; otherwise, it's negative
- The larger the number of Fiscal Year is, the less probability that a project would re-enter the funding pool
- The More Concurrent Projects there are, the higher probability there would be for projects that have had gaps to get re-funded

Using these predictors...

NIH can detect which investigators may be at higher risk of losing funding for their projects as well as design interventions to keep investigators funded, for example, providing supporting funds during gaps.

Future Directions

Estimating application volume in the future, based on current funded projects and the percentage of the all those expired projects that returned historically.

Identifying a reasonable time range for inspection. Of those projects that successfully get funded, how many may drop off in different periods of time and how far to look forward.

Providing insights for generating new predictors with complementary data that only NIH have access to, e.g. trial-and-fail funding application data.

Tagging "High-Risk Group" with more private data about investigators to better classify them. Features such as gender, race, age...

Thanks!

&

Questions?

Appendix

Project.Number	Type	Activity	IC	Serial.Number	Support.Year	Suffix	Program.Official.Information	Project.Start.Date	Project.End.Date	Study
5R37GM012633-34	5	R37	GM	12633	34		Unavailable	1977-12-01	1997-11-30	Speci
5R37GM012633-33	5	R37	GM	12633	33		Unavailable	1977-12-01	1997-11-30	Speci
5R37GM012633-32	5	R37	GM	12633	32		Unavailable	1977-12-01	1997-11-30	Speci
5R37GM012633-31	5	R37	GM	12633	31		Unavailable	1977-12-01	1997-11-30	Speci
4R37GM012633-30	4	R37	GM	12633	30		Unavailable	1977-12-01	1997-11-30	Speci
5R37GM032637-14	5	R37	GM	32637	14		Unavailable	1983-03-01	1996-11-30	Speci
5R37GM032637-13	5	R37	GM	32637	13		Unavailable	1983-03-01	1996-11-30	Speci
5R37GM032637-12	5	R37	GM	32637	12		Unavailable	1983-03-01	1996-11-30	Speci
5R37GM032637-11	5	R37	GM	32637	11		Unavailable	1983-03-01	1996-11-30	Speci
5R37GM024486-20	5	R37	GM	24486	20		Unavailable	1989-09-01	1997-11-30	Speci
5R37GM024486-19	5	R37	GM	24486	19		Unavailable	1989-09-01	1997-08-31	Speci
5R37GM024486-18	5	R37	GM	24486	18		Unavailable	1989-09-01	1997-08-31	Speci
5R37GM024486-17	5	R37	GM	24486	17		Unavailable	1989-09-01	1997-08-31	Speci
5R37GM035072-20	5	R37	GM	35072	20		RHOADES, MARCUS M.	1985-07-01	2005-12-02	Speci
5R37GM035072-19	5	R37	GM	35072	19		WOLFE, PAUL B.	1985-07-01	2005-06-30	Speci
5R37GM035072-18	5	R37	GM	35072	18		WOLFE, PAUL B.	1985-07-01	2005-06-30	Speci
5R37GM035072-17	5	R37	GM	35072	17		WOLFE, PAUL B.	1985-07-01	2005-06-30	Speci
4R37GM035072-16	4	R37	GM	35072	16		WOLFE, PAUL B.	1985-07-01	2005-06-30	Speci
5R37GM035072-15	5	R37	GM	35072	15		Unavailable	1985-07-01	2000-06-30	Virol

Showing 1 to 20 of 58,822 entries

Metadata:

NIGMS_R01_R37_93_08 : 58822 x 47

Appendix

PPID	Project.Number	Type	FY	Project.Start.Date	Project.End.Date	Budget.Start.Date	Budget.End.Date	Support.Year	Organization.State
1840203	1R01GM053905-01	1	1995	1995-05-17	1998-03-31	1995-05-17	1996-04-30	1	MA
1840203	2R01GM053905-04	2	1998	1995-05-17	2003-04-30	1998-05-01	1999-04-30	4	MA
1840203	1R01GM055781-01A2	1	1998	1998-09-15	2001-08-01	1998-09-15	1999-08-31	1	MA
1840203	1R01GM081336-01A1	1	2008	2008-09-01	2012-08-31	2008-09-01	2009-08-31	1	MA
1852586	2R01GM017980-29	2	1998	1978-09-01	2002-08-31	1998-09-01	1999-08-31	29	MA
1852586	2R01GM017980-33	2	2002	1978-09-01	2006-08-31	2002-09-01	2003-08-31	33	MA
1852586	2R01GM017980-37	2	2006	1978-09-01	2010-08-31	2006-09-01	2007-08-31	37	MA
1852587	2R01GM032134-17	2	1998	1983-01-01	2002-08-31	1998-09-01	1999-08-31	17	MA
1852587	2R01GM032134-21	2	2002	1983-01-01	2006-08-31	2002-09-04	2003-08-31	21	MA
1852587	2R01GM032134-25	2	2006	1983-01-01	2010-08-31	2006-09-01	2007-08-31	25	MA
1852587	2R01GM032134-29	2	2010	1983-01-01	2014-08-31	2010-09-06	2011-08-31	29	MA
1852587	1R01GM065519-01	1	2002	2002-04-01	2006-03-31	2002-04-01	2003-03-31	1	MA
1852587	2R01GM065519-05	2	2006	2002-04-01	2010-03-31	2006-04-01	2007-03-31	5	MA
1852587	2R01GM065519-09	2	2010	2002-04-01	2014-03-31	2010-04-01	2011-03-31	9	MA
1852587	2R01GM065519-13	2	2014	2002-04-01	2018-03-31	2014-04-01	2015-03-31	13	MA
1852589	2R01GM053567-06	2	2000	1995-09-30	2004-08-31	2000-09-01	2001-08-31	6	CT
1852589	1R01GM076661-01A2	1	2007	2007-04-15	2011-03-31	2007-04-15	2008-03-31	1	CT
1852589	2R01GM076661-05	2	2011	2007-04-15	2015-08-31	2011-09-30	2012-08-31	5	CT
1852589	2R01GM076661-09	2	2015	2007-04-15	2019-08-31	2015-09-01	2016-08-31	9	CT

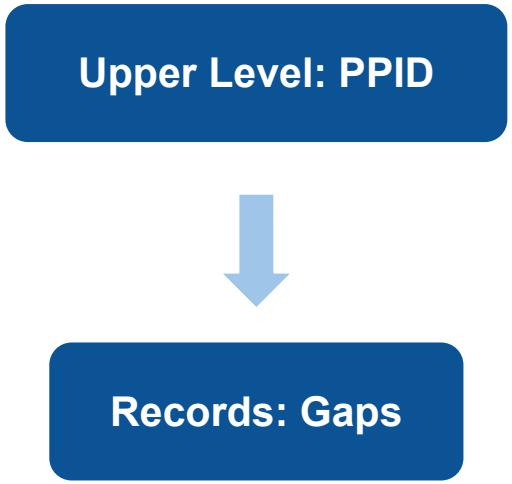
Showing 1 to 20 of 19,843 entries

Metadata:

IGMS_R01_R37_09_15 : 19843 x 47

Model Construction

- Model Selection: **Multilevel Logistic Model ➡ Logistic Model**



Reasons for using logistic regression finally:

Logistic regression model is more stable and doesn't have to adjust its parameters according to new sample before making predictions.

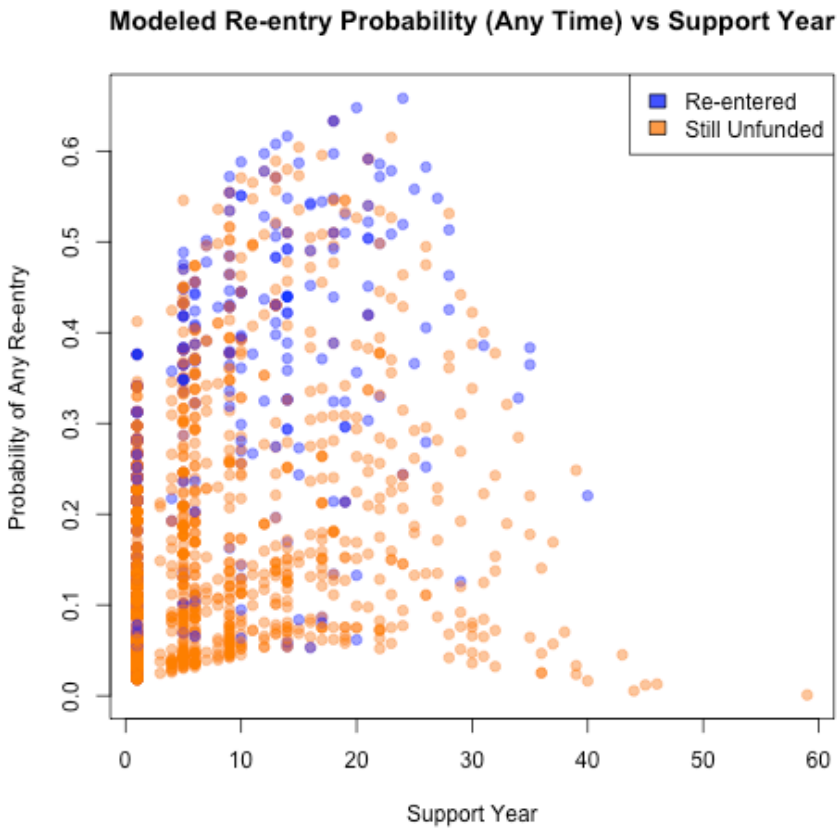
Since very few investigators have more than one projects, so the effect of their characteristic can be ignored.

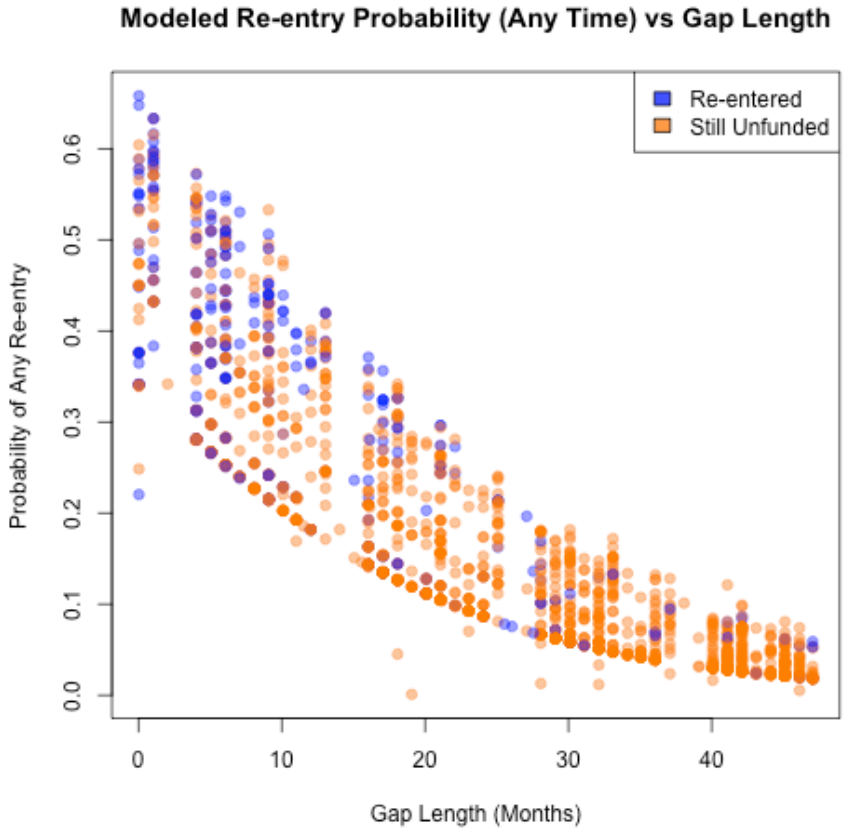
Appendix

Within different months, when controlling for other predictors, **1unit increase** in predictors will **lead the odds of re-entry**:

Predictors	% of change in Odds of re-entry (months)				
	6	12	24	36	Any
Gap Length	-7.937	-8.556	-7.754	-7.179	-6.837
# of Concurrent Project	21.593	23.759	11.902	10.063	16.442
Fiscal Year	-5.148	-3.422	-2.709	-2.360	-0.248
Support Year	8.829	9.615	10.447	11.761	14.643
$(\text{Support Year})^2$	-0.261	-0.288	-0.313	-0.345	-0.370

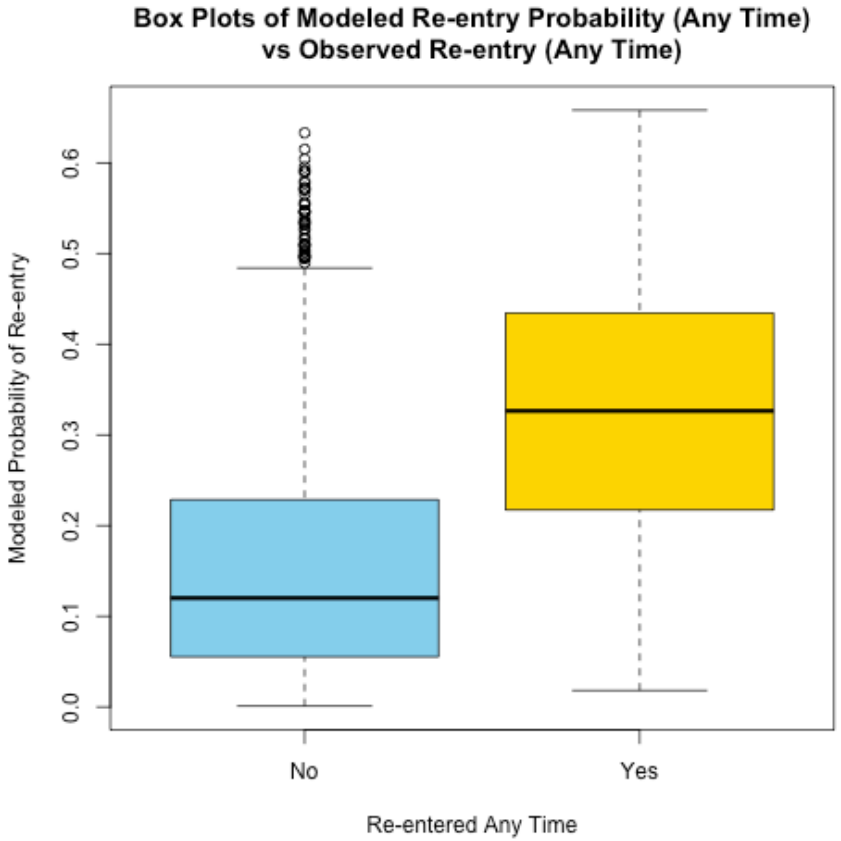
Appendix


Modeled Re-entry Probability (Any Time) vs Support Year
Re-entered
Still Unfunded
Probability of Any Re-entry
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
50
60
Support Year


Modeled Re-entry Probability (Any Time) vs Gap Length
Re-entered
Still Unfunded
Probability of Any Re-entry
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
Gap Length (Months)

Appendix

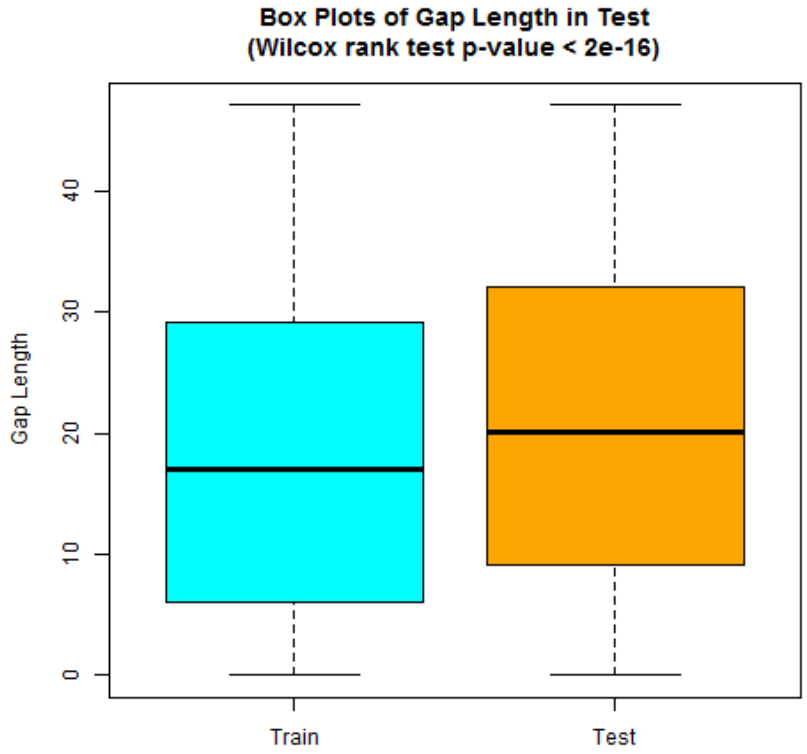
Model Validation - Using 2010 data



- The probability model predicted for project actually re-entered is higher than those for project didn't re-entered

Appendix

Model Validation - Using 2010 data



- General Gap Length in the training dataset (1998,2002, 2006) is different from in test dataset (2010).
- Gap Length has increased from 1998 to 2010.