Writing statistical models

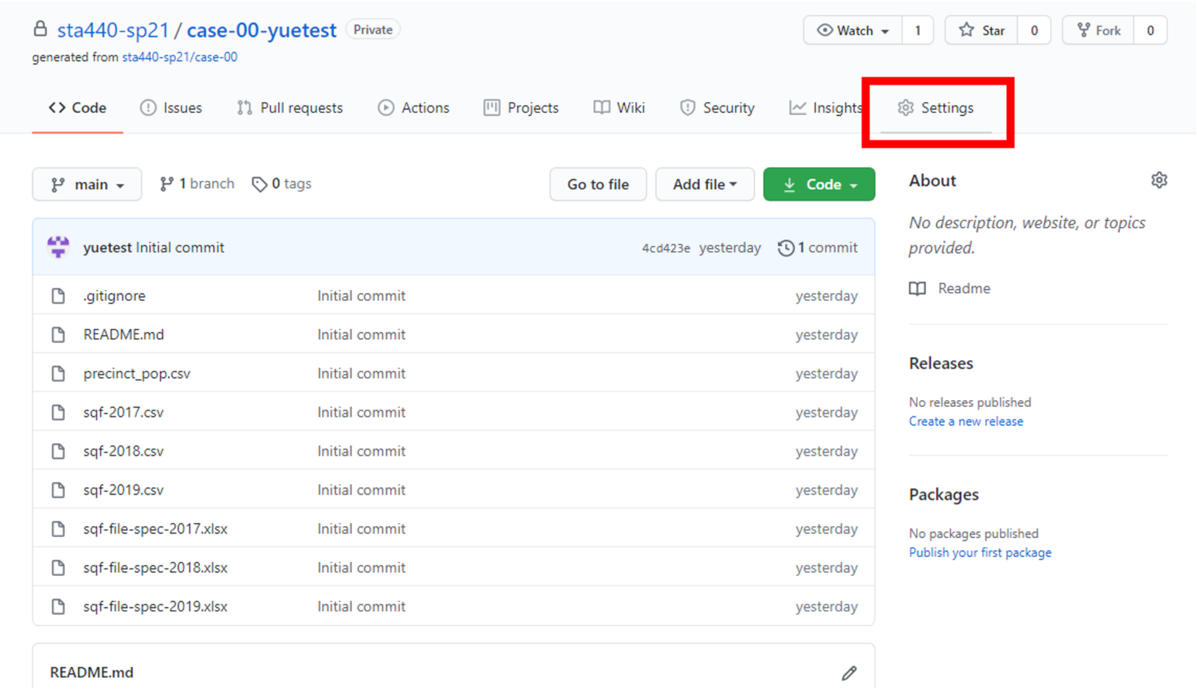
Yue Jiang

Duke University

A disclaimer

The following material was used during a live lecture. Without the accompanying oral comments and discussion, the text is incomplete as a record of the presentation. A full recording may be found via Zoom on the course Sakai site.

GitHub repositories



sta440-sp21 / case-00-yuetest Private

generated from sta440-sp21/case-00

yuetest Initial commit 4cd423e yesterday 1 commit

.gitignore	Initial commit	yesterday
README.md	Initial commit	yesterday
precinct_pop.csv	Initial commit	yesterday
sqf-2017.csv	Initial commit	yesterday
sqf-2018.csv	Initial commit	yesterday
sqf-2019.csv	Initial commit	yesterday
sqf-file-spec-2017.xlsx	Initial commit	yesterday
sqf-file-spec-2018.xlsx	Initial commit	yesterday
sqf-file-spec-2019.xlsx	Initial commit	yesterday

README.md

About

No description, website, or topics provided.

Readme

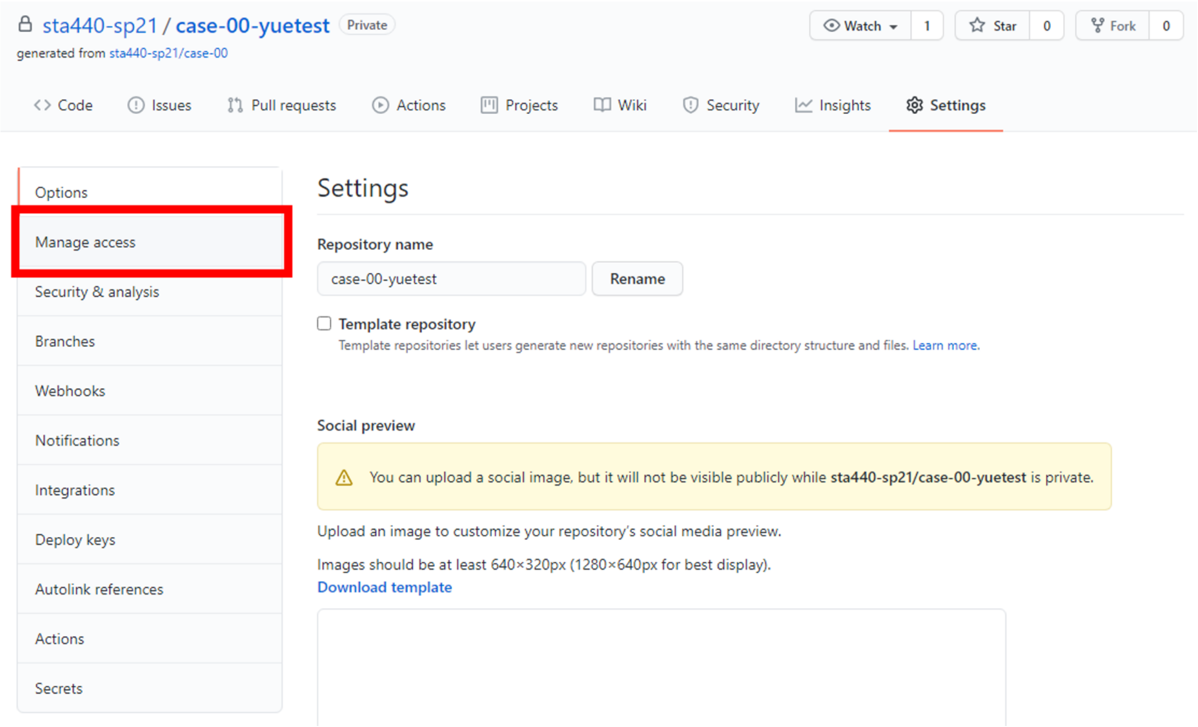
Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

GitHub repositories



Settings

Repository name

case-00-yuetest

□ Template repository

Template repositories let users generate new repositories with the same directory structure and files. Learn more.

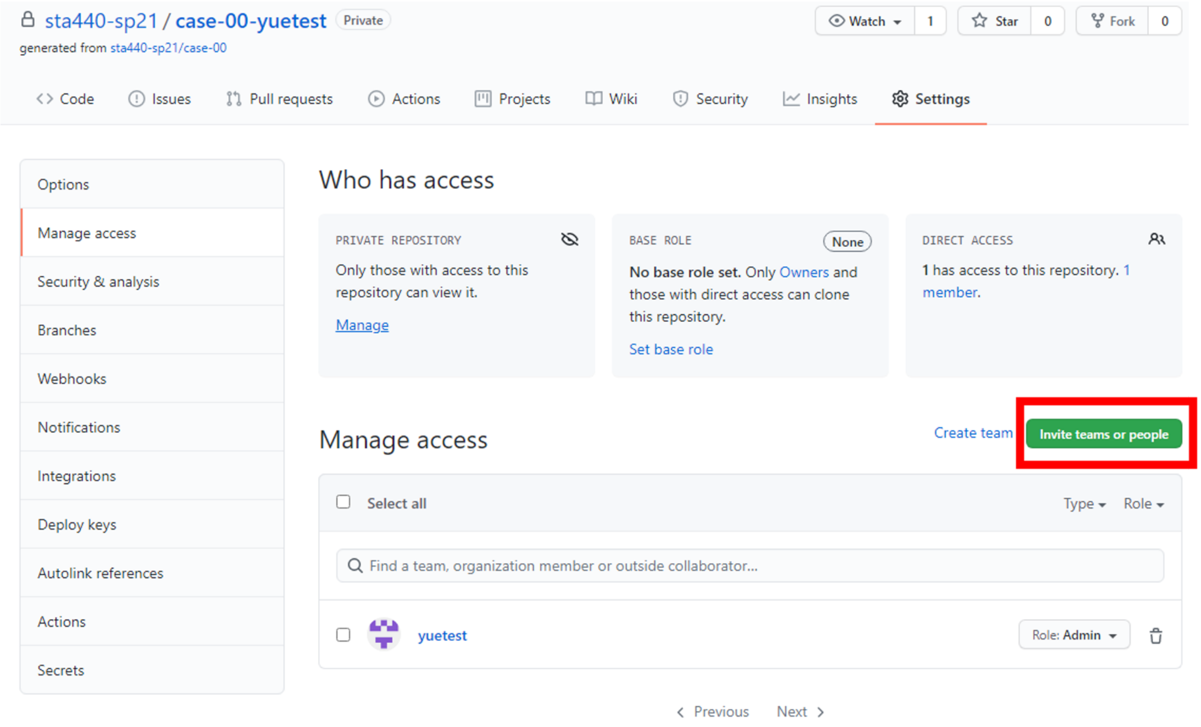
Social preview

You can upload a social image, but it will not be visible publicly while **sta440-sp21/case-00-yuetest** is private.

Upload an image to customize your repository's social media preview.

Images should be at least 640×320px (1280×640px for best display).

GitHub repositories



sta440-sp21 / case-00-yuetest Private

generated from sta440-sp21/case-00

Who has access

PRIVATE REPOSITORY

Only those with access to this repository can view it.

Manage

BASE ROLE None

No base role set. Only Owners and those with direct access can clone this repository.

Set base role

DIRECT ACCESS

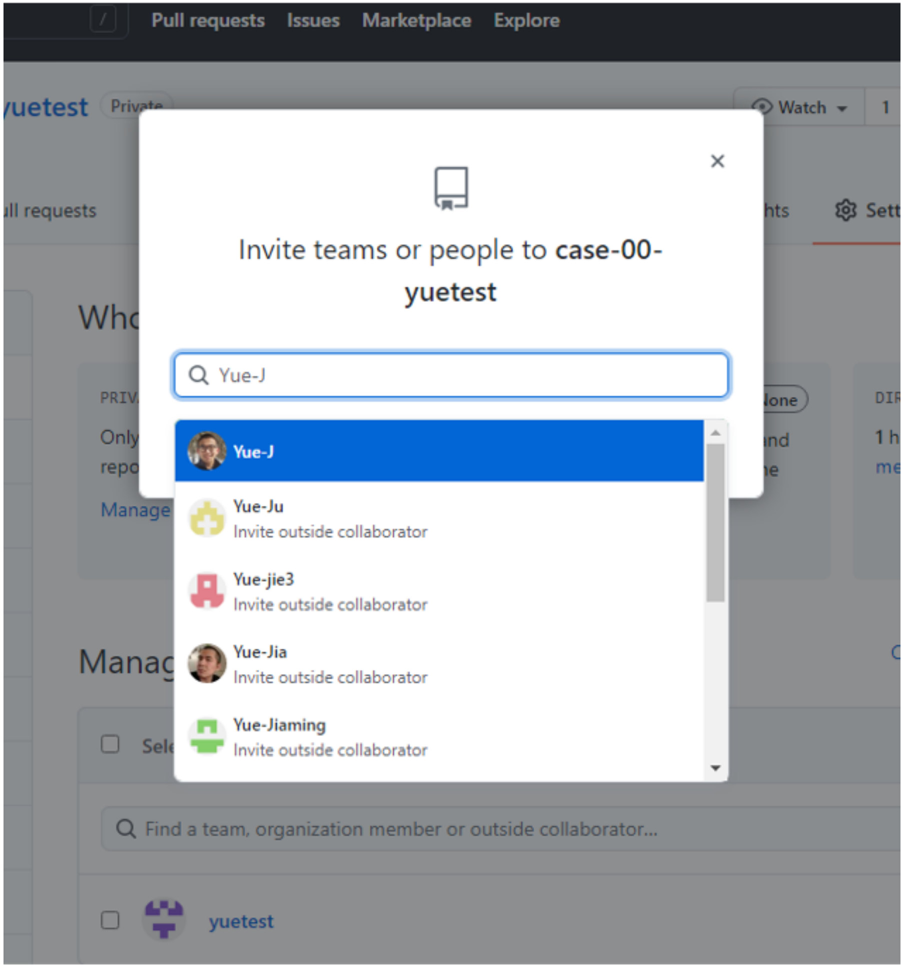
1 has access to this repository. 1 member.

Manage access

□ Select all

□ yuetest Role: Admin

GitHub repositories



Invite teams or people to **case-00-yuetest**

Yue-J

- Yue-J
- Yue-Ju
  Invite outside collaborator
- Yue-jie3
  Invite outside collaborator
- Yue-Jia
  Invite outside collaborator
- Yue-Jiaming
  Invite outside collaborator

Find a team, organization member or outside collaborator...

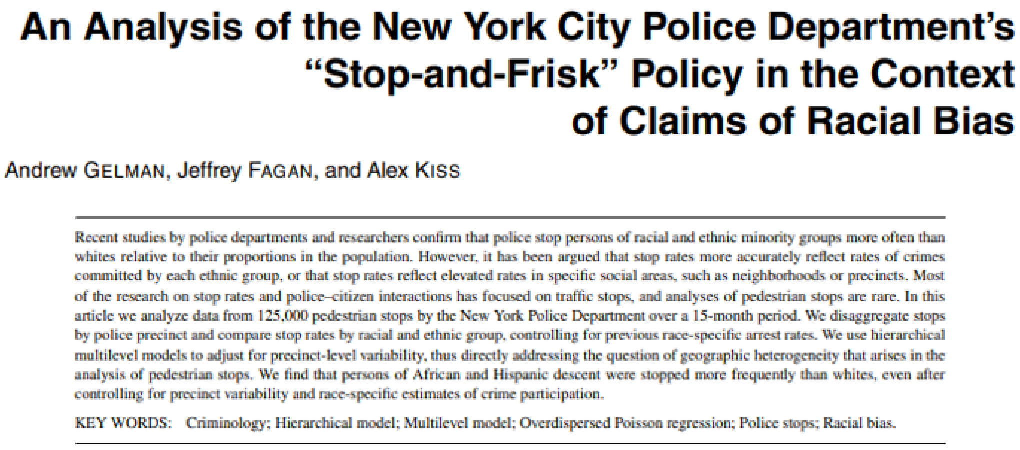
GitHub repositories

Stop-question-frisk

Photo - Adapted from Kena Betancur, Agence France-Presse/Getty Images

- Police protocol intended to reduce crime by stopping, questioning, and searching civilians
- Instituted in early 2000s, peaking in 2011 with almost 700,000 stops
- Fraught with controversy - many assert unfair targeting of Black and Hispanic citizens

Stop-question-frisk



An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias

Andrew GELMAN, Jeffrey FAGAN, and Alex KISS

Recent studies by police departments and researchers confirm that police stop persons of racial and ethnic minority groups more often than whites relative to their proportions in the population. However, it has been argued that stop rates more accurately reflect rates of crimes committed by each ethnic group, or that stop rates reflect elevated rates in specific social areas, such as neighborhoods or precincts. Most of the research on stop rates and police–citizen interactions has focused on traffic stops, and analyses of pedestrian stops are rare. In this article we analyze data from 125,000 pedestrian stops by the New York Police Department over a 15-month period. We disaggregate stops by police precinct and compare stop rates by racial and ethnic group, controlling for previous race-specific arrest rates. We use hierarchical multilevel models to adjust for precinct-level variability, thus directly addressing the question of geographic heterogeneity that arises in the analysis of pedestrian stops. We find that persons of African and Hispanic descent were stopped more frequently than whites, even after controlling for precinct variability and race-specific estimates of crime participation.

KEY WORDS: Criminology; Hierarchical model; Multilevel model; Overdispersed Poisson regression; Police stops; Racial bias.

- Gelman, Fagan, and Kiss (JASA, 2007) found evidence of racial disparities, even after adjusting for potential confounders
- A 2013 class-action lawsuit determined SQF was being used unconstitutionally; number of events fell sharply, with only around 10,000 stops per year from 2016 onwards

Writing a model

```r
# Note: you'll have to do the data cleaning yourself...
# Keep it clear and reproducible!

sqf <- read_csv("dat/sqf_cleaned.csv")
```

```r
table(sqf$build_clean)
```

```
##
##   HEA   MED OTHER   THN
##  3085 12443  1354 16310
```

```r
table(sqf$any_weapon)
```

```
##
##     0     1
## 30096  3096
```

Writing a model

```
summary(sqf$stop_duration_min)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00    8.00   11.75   14.00  999.00
```

```
summary(sqf$suspect_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   19.00   25.00   28.48   35.00   99.00
```

Writing a model

Let's ignore model assumption issues and diagnostics for now. Suppose you wanted to create a linear model that related the stop duration of an SQF event to the build of the suspect, whether they had a weapon on their person, and their age. How would you express such a model mathematically?

What could go wrong?

What issue(s) do you see with the model below? (Again, ignore issues of whether the model itself is appropriate; assume we just want a linear model relating the outcome of stop duration to the three predictors)

$$duration = build + weapon + age$$

What could go wrong?

How about here?

$$duration = \beta_1 build + \beta_2 weapon + \beta_3 age$$

What could go wrong?

Or here?

$$\begin{aligned} duration = {} & \beta_1 I(build = HEA) + \beta_2 I(build = MED) + \\ & \beta_3 I(build = OTHER) + \beta_4 weapon + \\ & \beta_5 age \end{aligned}$$

What could go wrong?

Are we finished?

$$\begin{aligned} duration_i = {} & \beta_0 + \beta_1 I(build_i = HEA) + \beta_2 I(build_i = MED) + \\ & \beta_3 I(build_i = OTHER) + \beta_4 weapon_i + \\ & \beta_5 age_i \end{aligned}$$

What could go wrong?

What are the parameters in the model written below? How can we write it succintly in matrix notation?

$$
\begin{aligned}
duration_i = \beta_0 + \beta_1 I(build_i = HEA) + \beta_2 I(build_i = MED) + \\
\beta_3 I(build_i = OTHER) + \beta_4 weapon_i + \\
\beta_5 age_i + \epsilon_i
\end{aligned}
$$

where

$$\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

A Bayesian model

What about a Bayesian alternative to this OLS model? How would you formulate it? What about priors on the parameters?

Another model

Say instead of duration being the outcome variable, suppose we're interested in the log-odds of having a "long" SQF event (an event lasting over half an hour). If we fit a logistic regression model for this question, what would our model look like?

```
sqf %>%
  mutate(long_sqf = ifelse(stop_duration_min > 30, 1, 0)) %>%
  count(long_sqf)
```

```
##   long_sqf     n
## 1        0 31731
## 2        1  1461
```

References

[1] Stop, Question, and Frisk Data, 2003 - 2019. Accessed on NYPD website.

[2] Gelman A, Fagan J, Kiss A (2012). "Stop-and-frisk policy in the context of claims of racial bias." Journal of the American Statistical Association. 102(479): 813 - 823.

[3] Floyd, et al. v. City of New York, et al. United States District Court for the Southern District of New York. 959 F. Supp. 2d 540 (2013).

[4] Mummolo, J (2018). "Modern Police Tactics, Police-Citizen Interactions, and the Prospects for Reform". The Journal of Politics. 80: 1–15.