ECON 251 Discussion

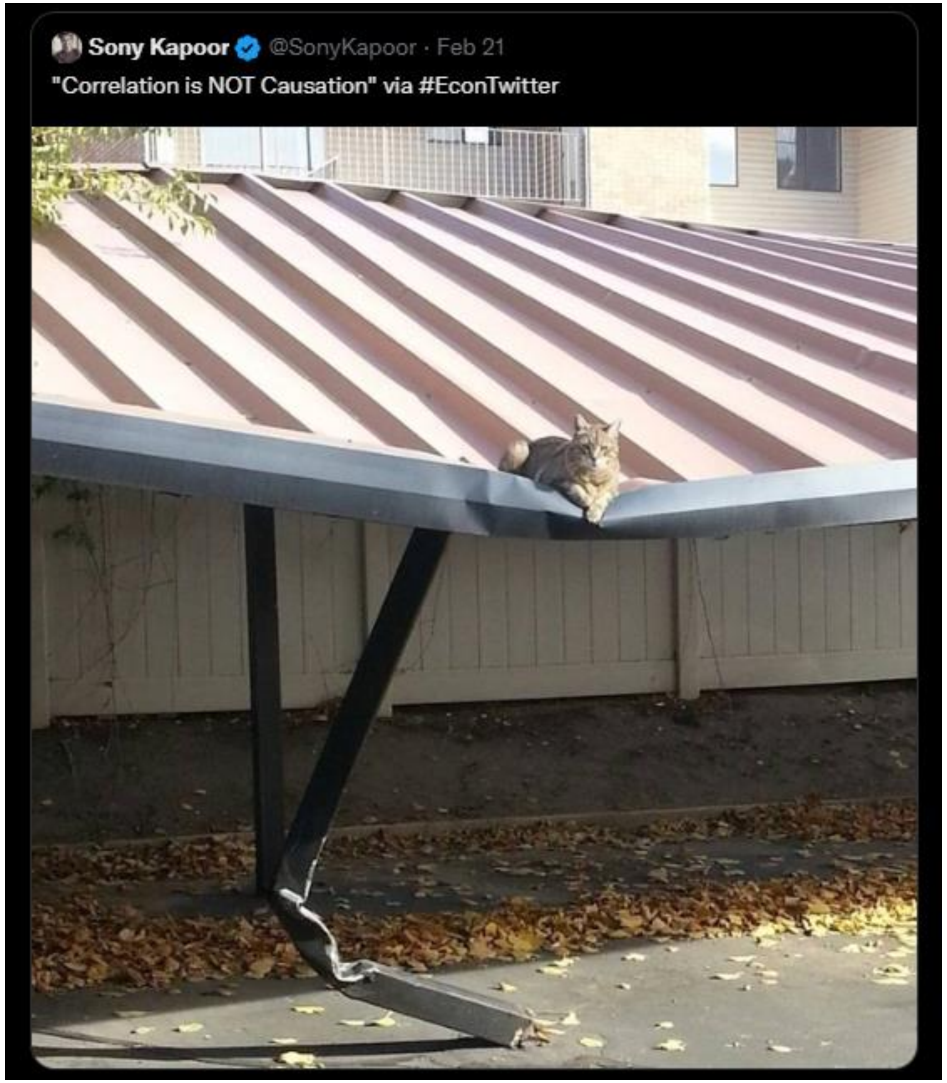
Linear Probability Model (LPM) + Diff-in-Diff (DD)

Elird Haxhiu

Fall 2022

Last discussion of the semester!

- Congratulations!
- Technically almost done…
- Still just the beginning
- Course evaluations


Sony Kapoor @SonyKapoor · Feb 21
"Correlation is NOT Causation" via #EconTwitter

Tanya Cashorali
@tanyacash21

We are living in the future. Holy shit. #rstats #rvest #ChatGPT

Help

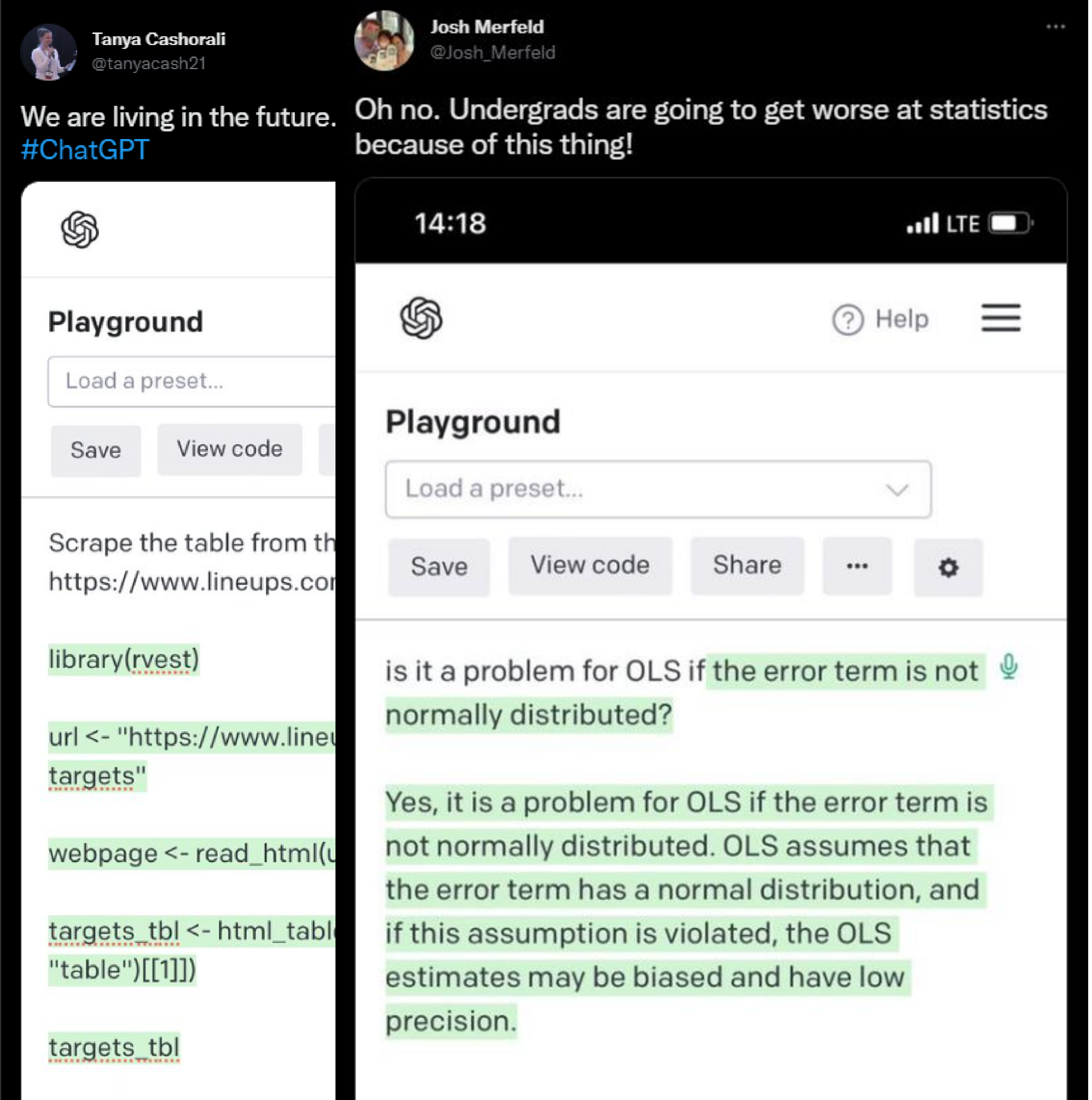
Playground

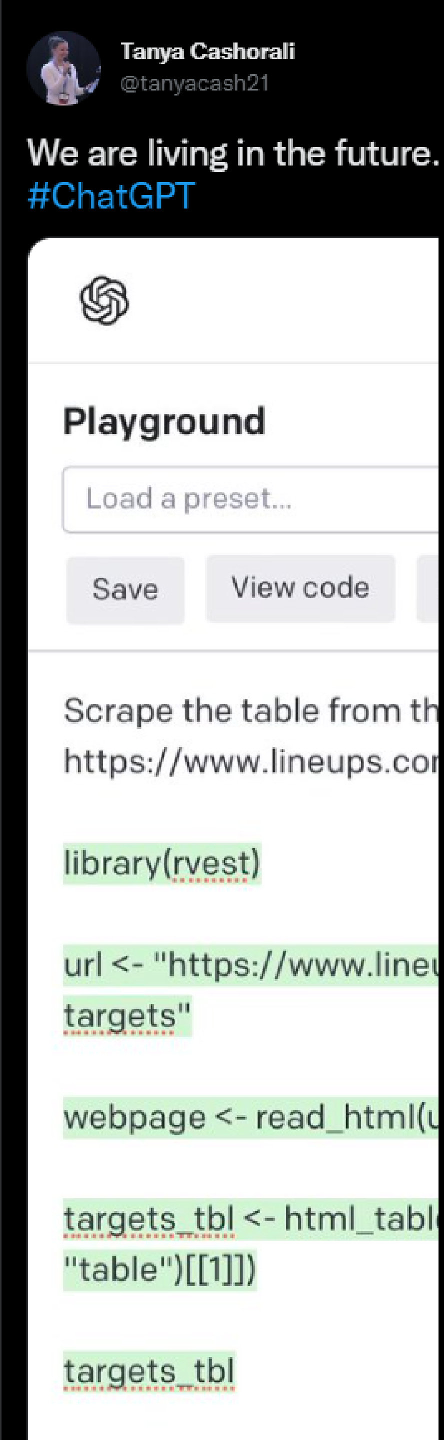
Load a preset...

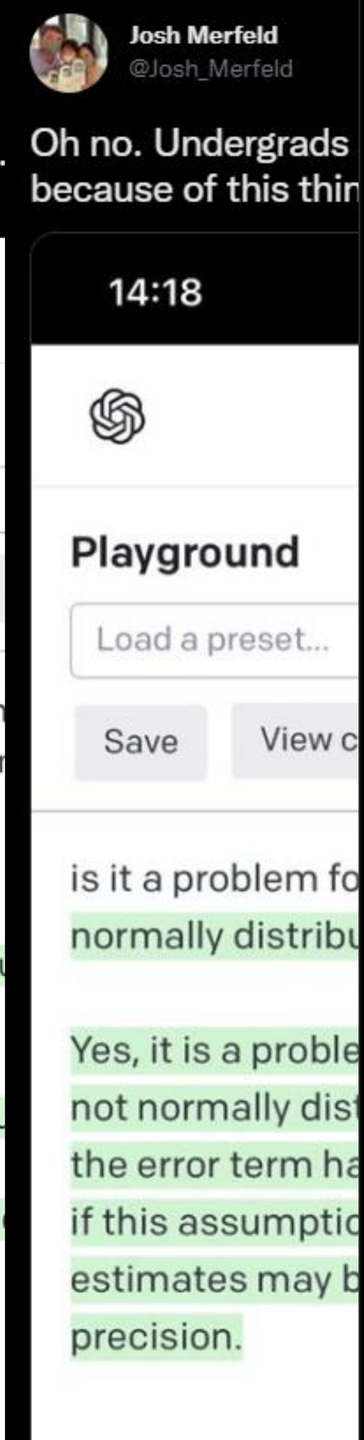
Save | View code | Share | ···

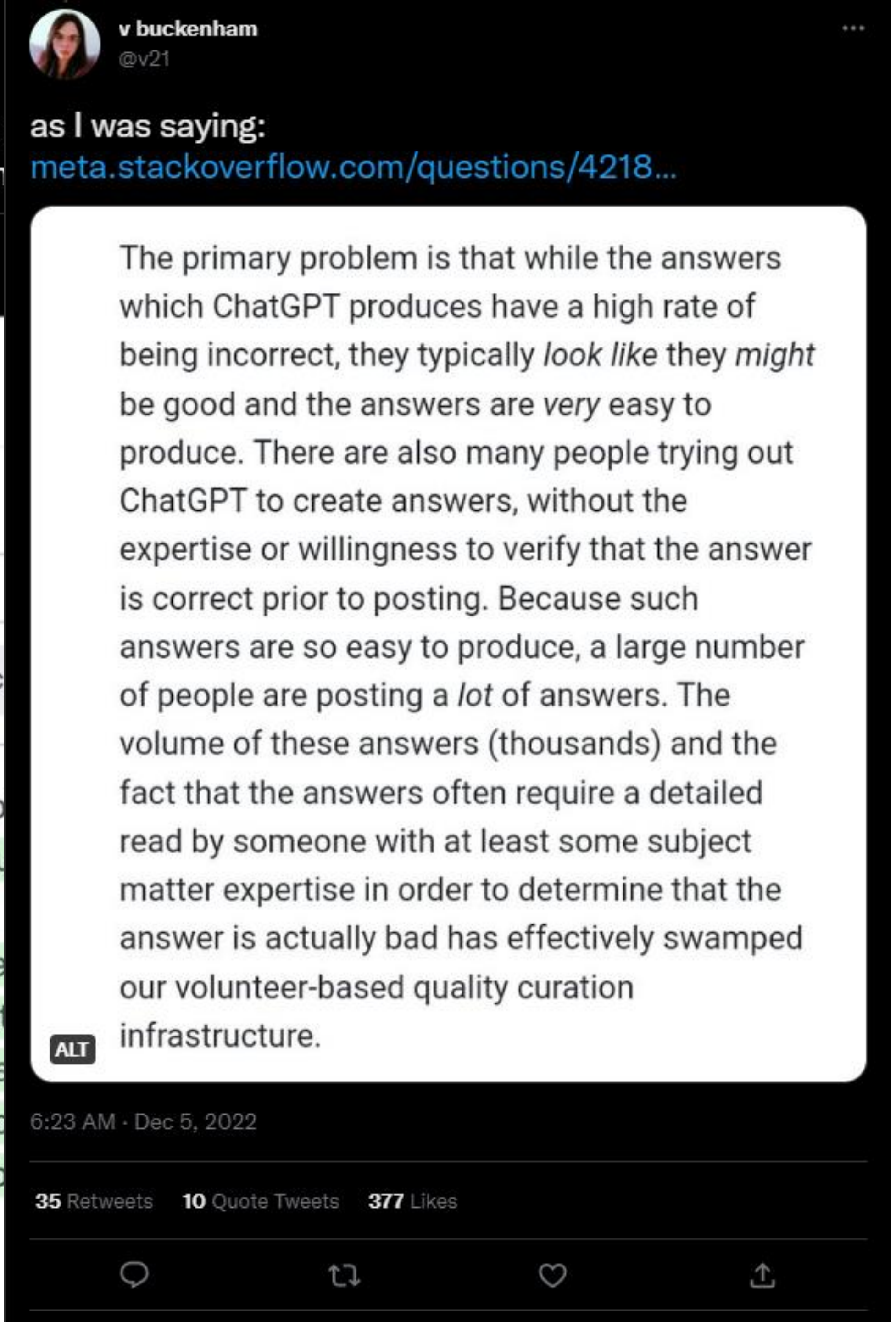
Scrape the table from this website in R https://www.lineups.com/nfl/nfl-targets

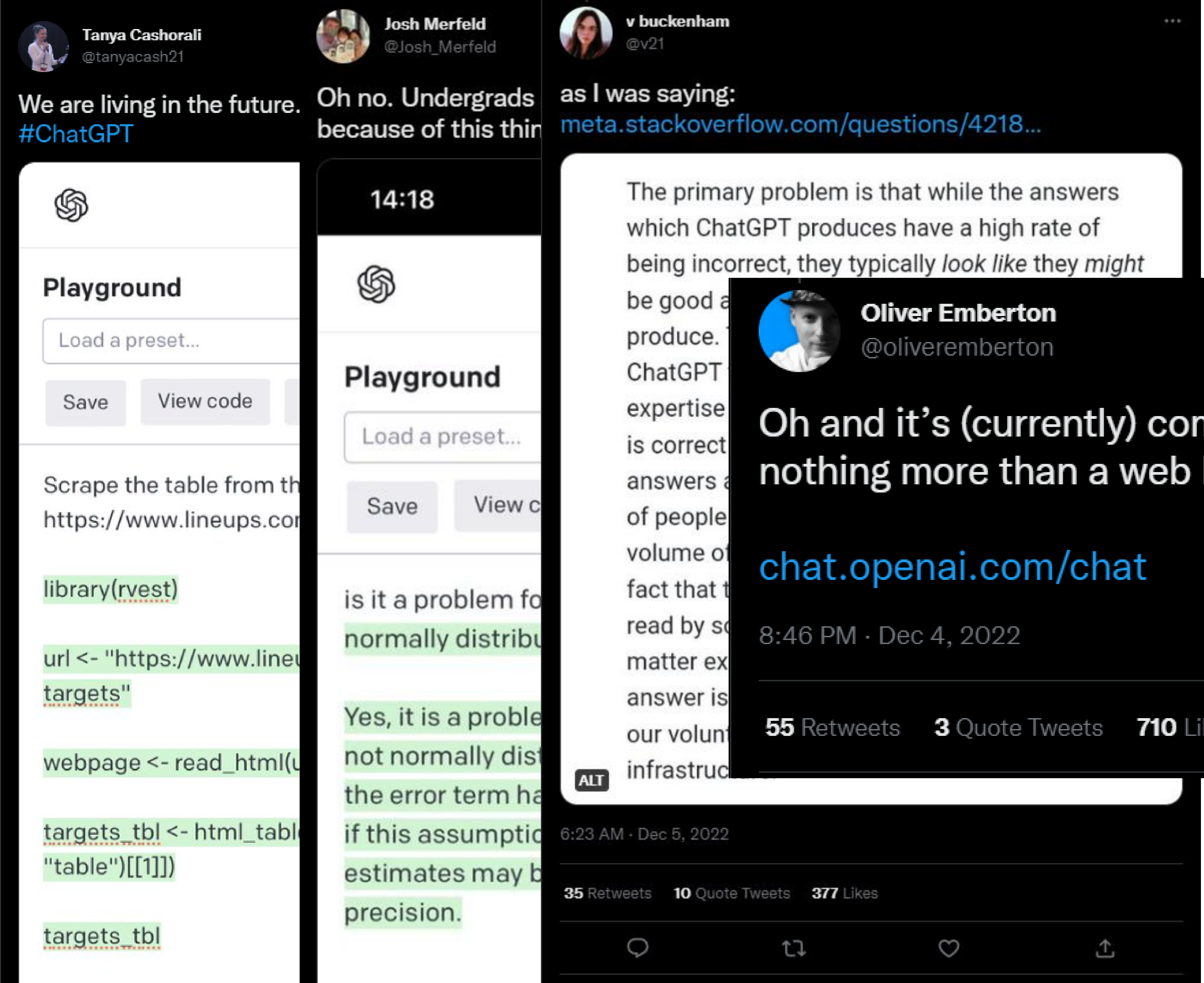
```
library(rvest)

url <- "https://www.lineups.com/nfl/nfl-targets"

webpage <- read_html(url)

targets_tbl <- html_table(html_nodes(webpage, "table")[[1]])

targets_tbl
```

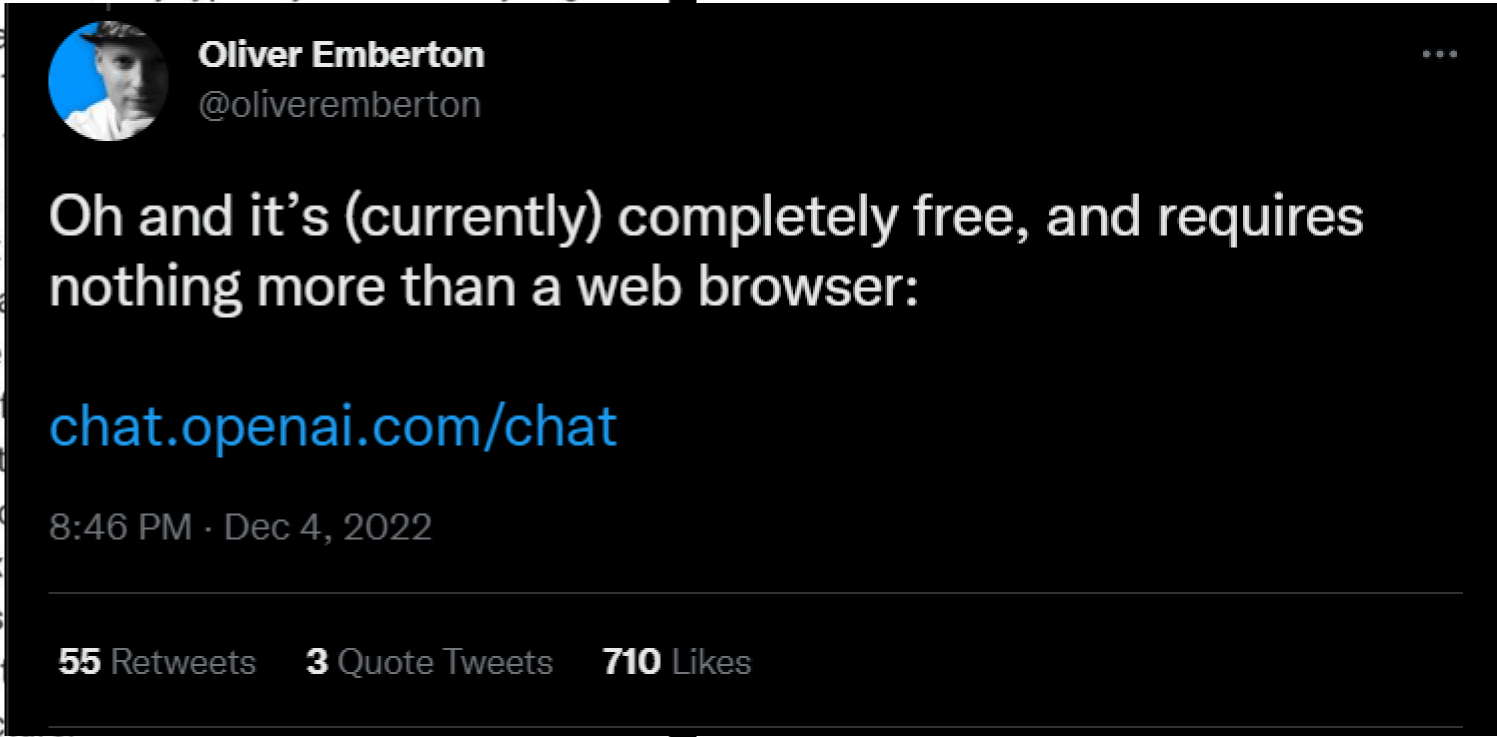

Tanya Cashorali
@tanyacash21
We are living in the future. #ChatGPT
Playground
Load a preset...
Save
View code
Scrape the table from th
https://www.lineups.cor
library(rvest)
url <- "https://www.lineu
targets"
webpage <- read_html(u
targets_tbl <- html_tabl
"table")[[1]])
targets_tbl
Josh Merfeld
@Josh_Merfeld
Oh no. Undergrads are going to get worse at statistics because of this thing!
14:18
LTE
Help
Playground
Load a preset...
Save
View code
Share
is it a problem for OLS if the error term is not normally distributed?
Yes, it is a problem for OLS if the error term is not normally distributed. OLS assumes that the error term has a normal distribution, and if this assumption is violated, the OLS estimates may be biased and have low precision.


Tanya Cashorali
@tanyacash21
We are living in the future. #ChatGPT
Playground
Load a preset...
Save
View code
Scrape the table from th
https://www.lineups.co
library(rvest)
url <- "https://www.line
targets"
webpage <- read_html(
targets_tbl <- html_tabl
"table")[[1]])
targets_tbl


Josh Merfeld
@Josh_Merfeld
Oh no. Undergrads
because of this thi
14:18
Playground
Load a preset...
Save
View c
is it a problem fo
normally distrib
Yes, it is a proble
not normally dis
the error term ha
if this assumpti
estimates may b
precision.


v buckenham
@v21
as I was saying:
meta.stackoverflow.com/questions/4218...
The primary problem is that while the answers which ChatGPT produces have a high rate of being incorrect, they typically *look like* they *might* be good and the answers are *very* easy to produce. There are also many people trying out ChatGPT to create answers, without the expertise or willingness to verify that the answer is correct prior to posting. Because such answers are so easy to produce, a large number of people are posting a *lot* of answers. The volume of these answers (thousands) and the fact that the answers often require a detailed read by someone with at least some subject matter expertise in order to determine that the answer is actually bad has effectively swamped our volunteer-based quality curation infrastructure.
ALT
6:23 AM · Dec 5, 2022
35 Retweets 10 Quote Tweets 377 Likes


Tanya Cashorali
@tanyacash21
We are living in the future.
#ChatGPT
Playground
Load a preset...
Save
View code
Scrape the table from th
https://www.lineups.cor
library(rvest)
url <- "https://www.lineu
targets"
webpage <- read_html(u
targets_tbl <- html_tabl
"table")[[1]])
targets_tbl
Josh Merfeld
@Josh_Merfeld
Oh no. Undergrads
because of this thin
14:18
Playground
Load a preset...
Save
View c
is it a problem fo
normally distribu
Yes, it is a proble
not normally dist
the error term ha
if this assumptic
estimates may b
precision.
v buckenham
@v21
as I was saying:
meta.stackoverflow.com/questions/4218...
The primary problem is that while the answers
which ChatGPT produces have a high rate of
being incorrect, they typically look like they might
be good a
produce.
ChatGPT
expertise
is correct
answers a
of people
volume of
fact that t
read by so
matter ex
answer is
our volunt
infrastruc
ALT
6:23 AM · Dec 5, 2022
35 Retweets 10 Quote Tweets 377 Likes


Oliver Emberton
@oliveremberton
Oh and it's (currently) completely free, and requires nothing more than a web browser:
chat.openai.com/chat
8:46 PM · Dec 4, 2022
55 Retweets 3 Quote Tweets 710 Likes

Outline

1. Linear probability model (LPM) for discrete outcomes
2. Review HW3 solutions
3. Difference-in-differences in practice

Continuous Outcomes $Y \in \mathbb{R}$

- Independence + continuous Y gives usual "slope interpretation"

$$Y = \beta_0 + \beta_1 X + U$$

$$\begin{aligned} E[Y|X] &= E[\beta_0 + \beta_1 X + U|X] \\ &= \beta_0 + \beta_1 X + E[U|X] \\ &= \beta_0 + \beta_1 X \end{aligned}$$

$$\Rightarrow \beta_1 = \frac{\partial}{\partial X} E[Y|X]$$

Binary Outcomes $Y \in \{0,1\}$

- Independence + binary Y gives "change in prob(Y=1)" interpretation

$$Y = \beta_0 + \beta_1 X + U$$

$$\begin{aligned} E[Y|X] &= P[Y = 1|X] \cdot 1 + P[Y = 1|X] \cdot 0 \\ &= P[Y = 1|X] \end{aligned}$$

$$\Rightarrow \beta_1 = \frac{\partial}{\partial X} P[Y = 1|X]$$

Linear Probability Model (LPM)

- OLS estimates of linear model with binary outcome

- LPM is nice because…
  1. Easy to estimate
  2. Easy to interpret

- LPM is problematic since
  1. Predicted values of outcome can be outside of $[0,1]$ interval
  2. Does not make sense for X to change $P[Y = 1|X]$ linearly

Linear Probability Model (LPM)

- LPM is problematic because…
  1. Predicted values of outcome can be outside of $[0,1]$ interval
  2. Does not make sense for X to change $P[Y = 1|X]$ linearly
  3. Homoskedasticity is always violated

$$\begin{aligned}
\mathrm{Var}(Y|X) &= E[Y^2|X] - E(Y|X)^2 \\
&= [P(Y = 1|X) \cdot 1^2 + P(Y = 0|X) \cdot 0^2] - [P(Y = 1|X)]^2 \\
&= P(Y = 1|X) - P(Y = 1|X)^2 \\
&= P(Y = 1|X)[1 - P(Y = 1|X)]
\end{aligned}$$

Alternative Estimators = assume $\boldsymbol{U}$ distribution

- Probit = assume that $U \sim N(0,1)$

$$P(Y=1|X) = \Phi(\beta_0 + \beta_1 X)$$

where $\Phi(u) \coloneqq P[U \leq u] = F_U(u)$ denotes the standard normal CDF

- Logit = assume that U follows logistic distribution with PDF

$$f_U(u) = \frac{1}{1+e^{-u}}$$

Review HW3 solutions

Outline

1. Linear probability model (LPM) for discrete outcomes
2. Review HW3 solutions
3. Difference-in-differences in practice

Difference-in-differences = compare Y change of units exposed to some policy T with Y change of unexposed

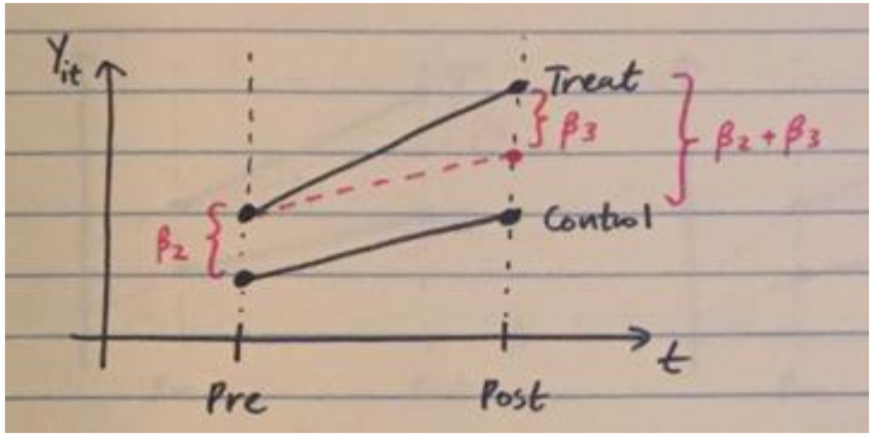
2 periods (before/after) and 2 groups (treated/control)

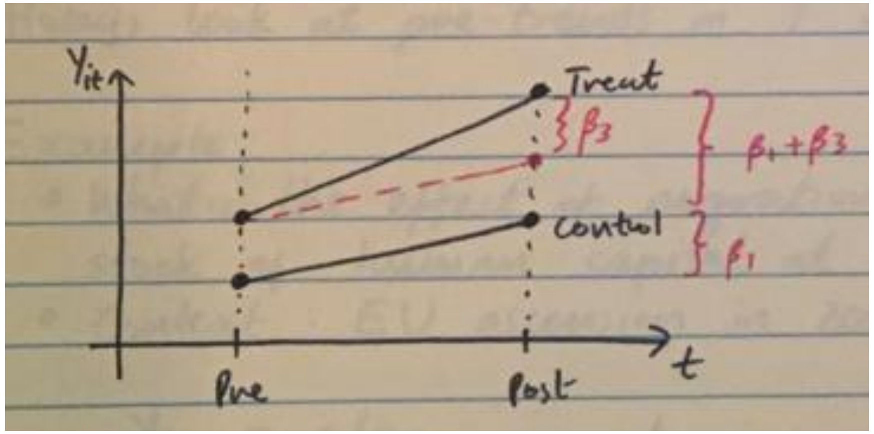
$Y_{it} :=$ outcome of interest

$P_t := 1\{t \text{ is after treatment occurs}\}$

$T_i := 1\{i \text{ is treated/exposed}\}$

$$Y_{it} = \beta_0 + \beta_1 P_t + \beta_2 T_i + \beta_3 [P_t \cdot T_i] + U_i$$

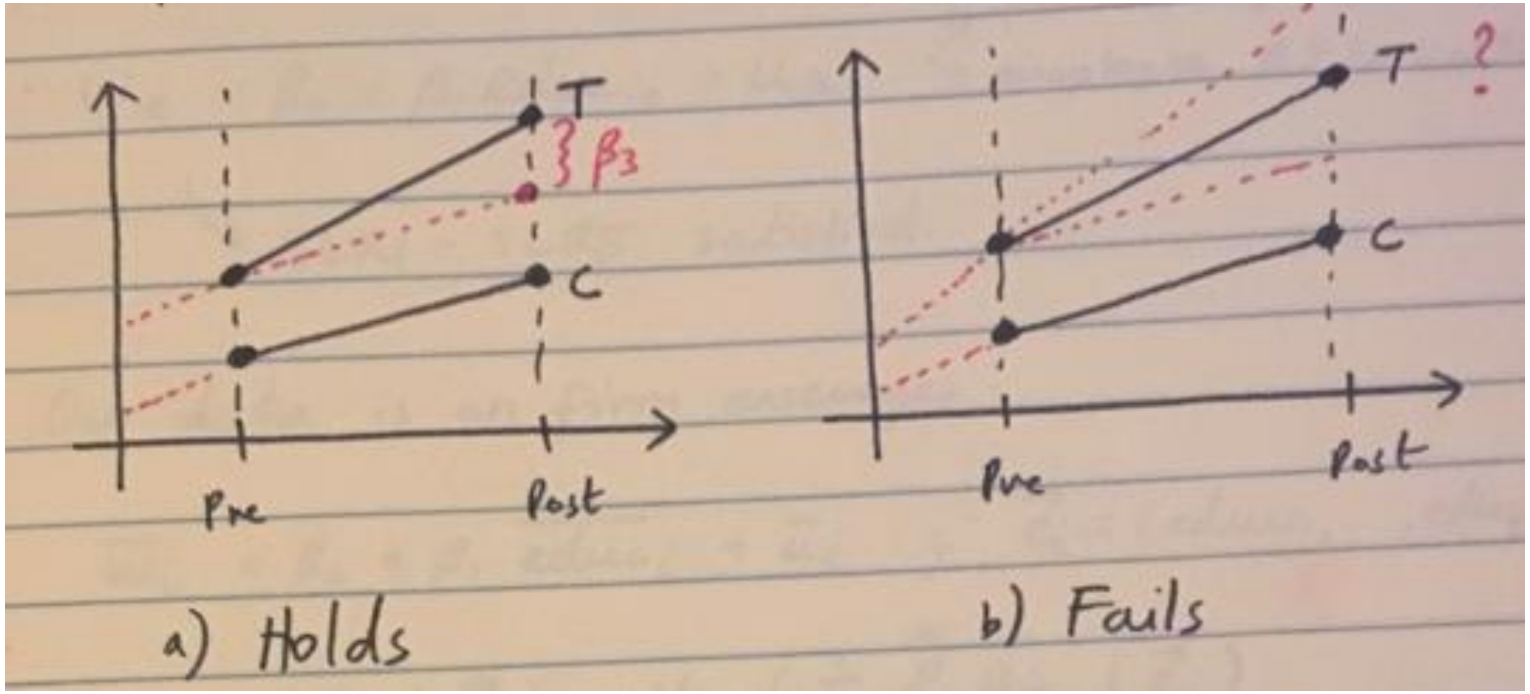




	Before	After	After – Before
Control	β_0	$\beta_0 + \beta_1$	β_1
Treated	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_1 + \beta_3$
Treat – Control	β_2	$\beta_2 + \beta_3$	β_3

Parallel Trends Assumption = exposed units Y without policy T would have changed like unexposed units Y

- PTA is an untestable assumption, just like OLS exogeneity or IV exogeneity
- However, if we have access to more data before policy, we can assess how likely it is to hold in practice… commonly known as "checking for pre-trends"
- One reason why people seem to like DD… visual check of identifying assumption!



Quasi-Market Competition in Public Service Provision: User Sorting and Cream-Skimming

Thorbjørn Sejr Guul[*,†], **Ulrik Hvidman**[†], **Hans Henrik Sievertsen**[‡]

[*]TrygFonden's Centre for Child Research; [†]Aarhus University; [‡]VIVE – The Danish Center for Social Science Research, IZA, University of Bristol

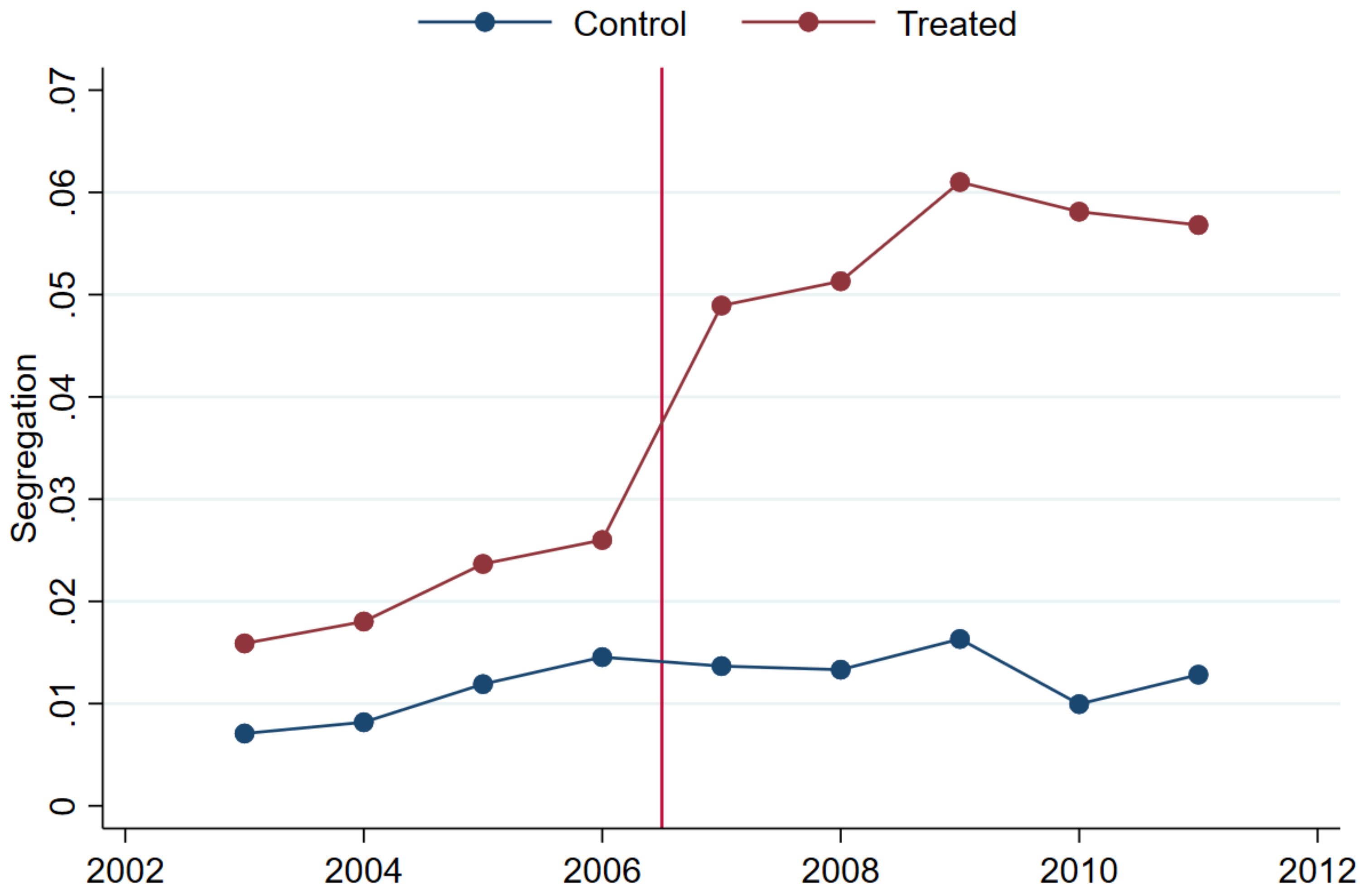
Address correspondence to the author at tsg@ps.au.dk.

Abstract

Quasi-markets that introduce choice and competition between public service providers are intended to improve quality and efficiency. This article demonstrates that quasi-market competition may also affect the distribution of users. First, we develop a simple theoretical framework that distinguishes between user sorting and cream-skimming as mechanisms through which quasi-markets may lead to high-ability users becoming more concentrated among one group of providers and low-ability users among a different group. Second, we empirically examine the impact of a nationwide quasi-market policy that introduced choice and activity-based budgeting into Danish public high schools. We exploit variation in the degree of competition that schools were exposed to, based on the concentration of providers within a geographical area. Using a differences-in-differences design—and register data containing the full population of students over a 9-year period ($N = 207{,}394$)—we show that the composition of students became more concentrated in terms of intake grade point average after the reform in high-competition areas relative to low-competition areas. These responses in high-competition regions appear to be driven both by changes in user sorting on the demand side and by cream-skimming behavior among public providers on the supply side.

In Stata!

Code from Hans Henrik Sivertsen



In Stata!

Code from Hans Henrik Sivertsen

```
. reg segregation Post Treated PostXTreated
```

Source	SS	df	MS
Model	.006163964	3	.002054655
Residual	.000222337	14	.000015881
Total	.006386301	17	.000375665

Number of obs	=	18
F(3, 14)	=	129.38
Prob > F	=	0.0000
R-squared	=	0.9652
Adj R-squared	=	0.9577
Root MSE	=	.00399

segregation	Coefficient	Std. err.	t	P>|t|
Post	.002788	.0026733	1.04	0.315
Treated	.010465	.0028179	3.71	0.002
PostXTreated	.031551	.0037806	8.35	0.000
_cons	.01043	.0019926	5.23	0.000