RELATIONSHIPS

MPA 635: Data Visualization

October 16, 2018

PLAN FOR TODAY

Dual y-axes

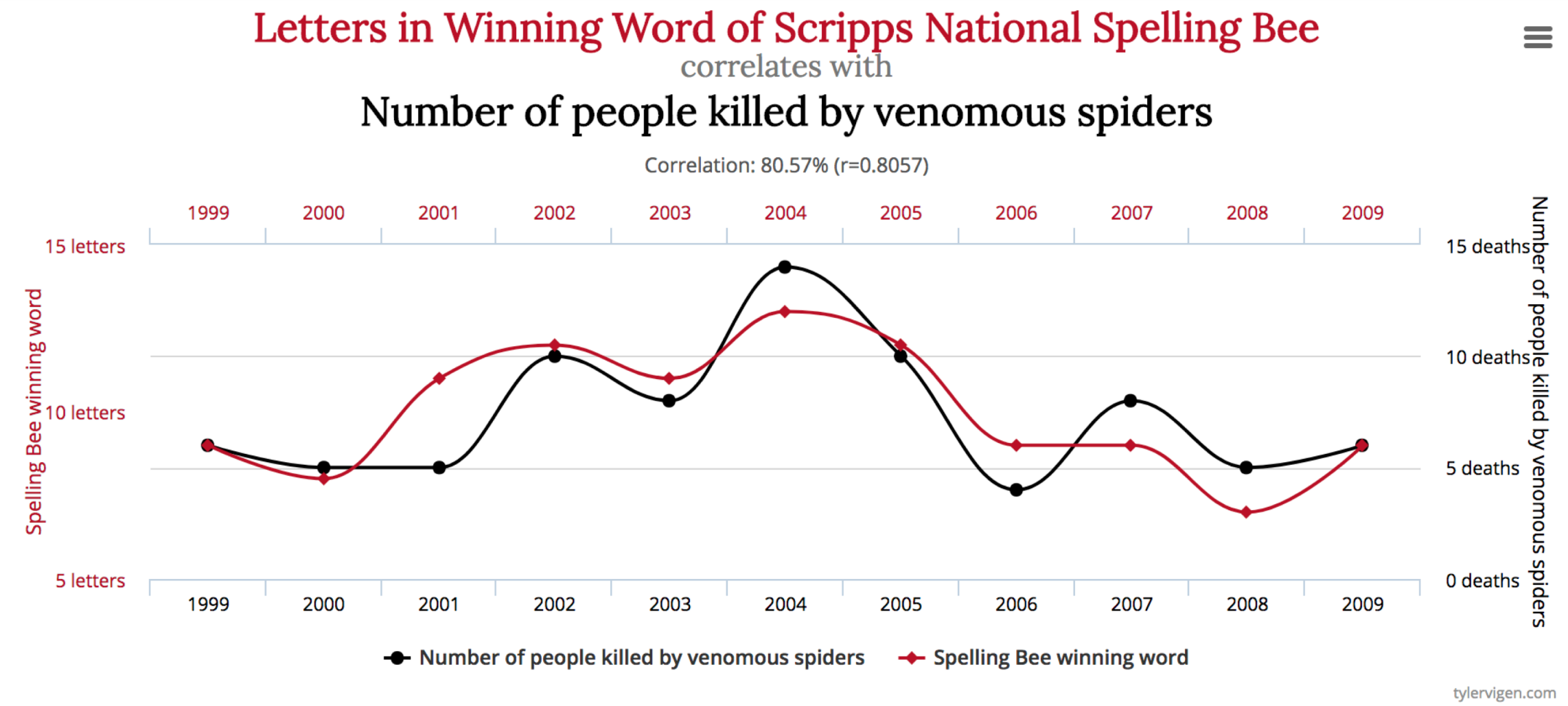
Correlation and regression

Correlation and causation

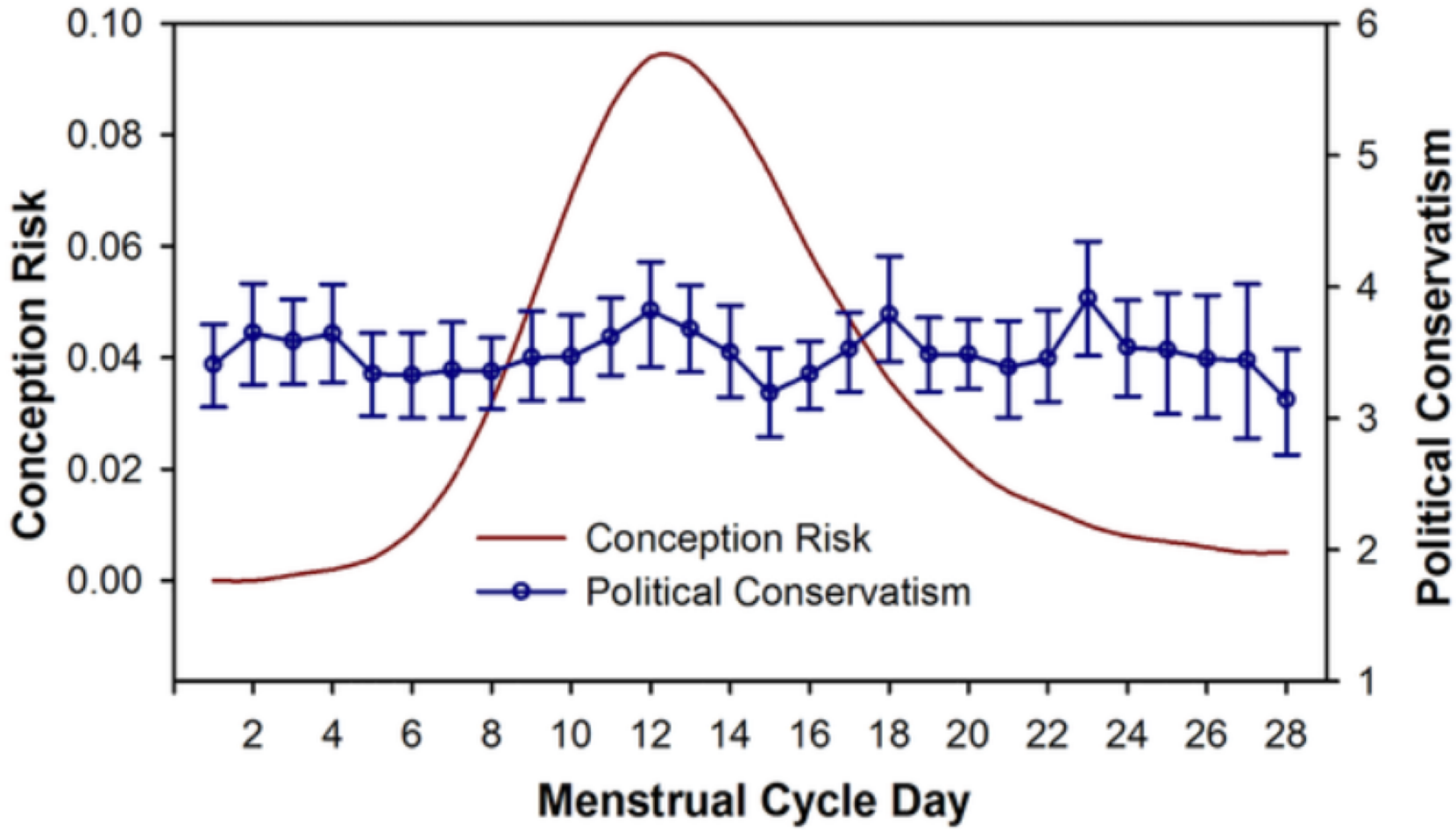
ggplot examples

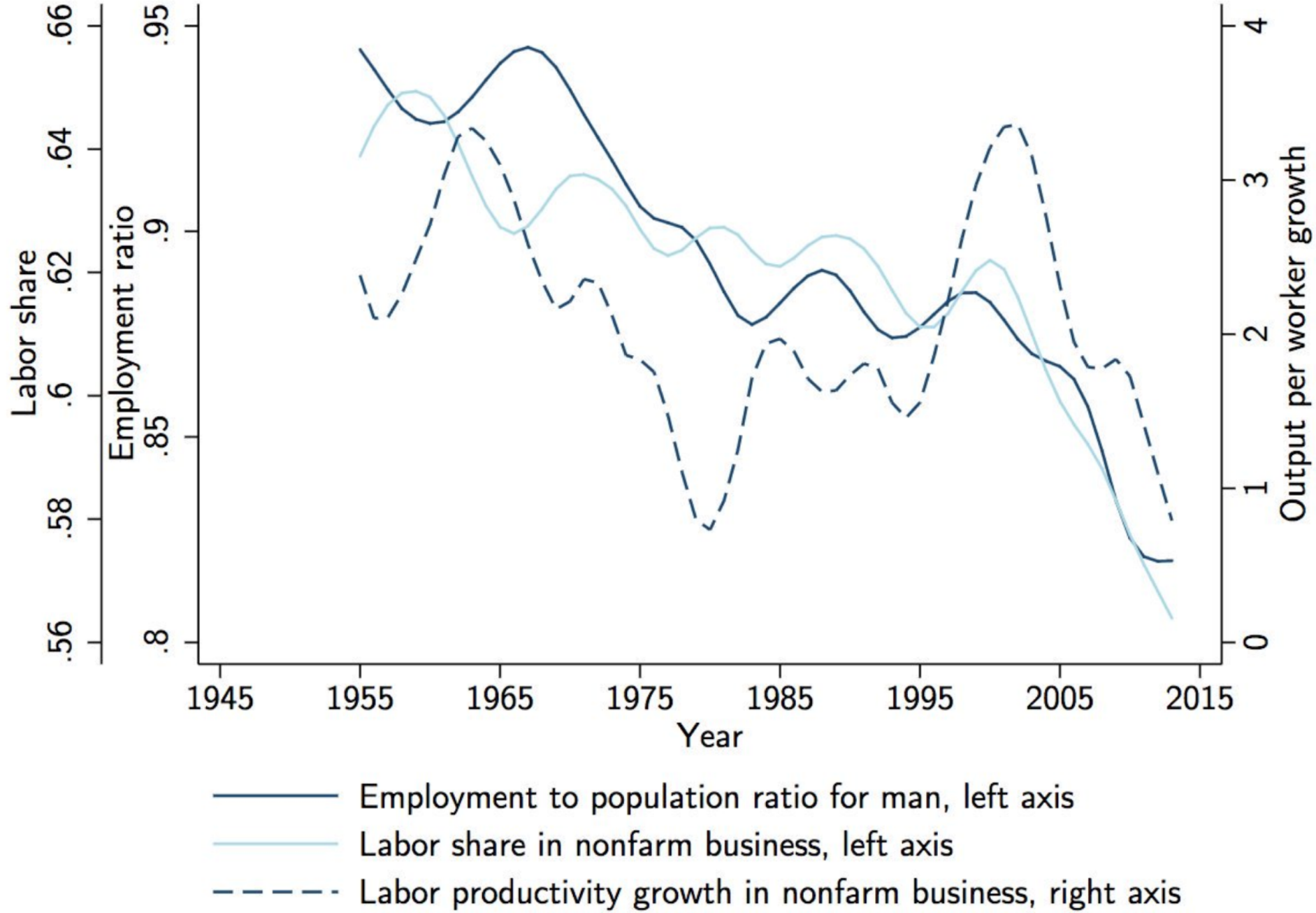
DUAL Y-AXES

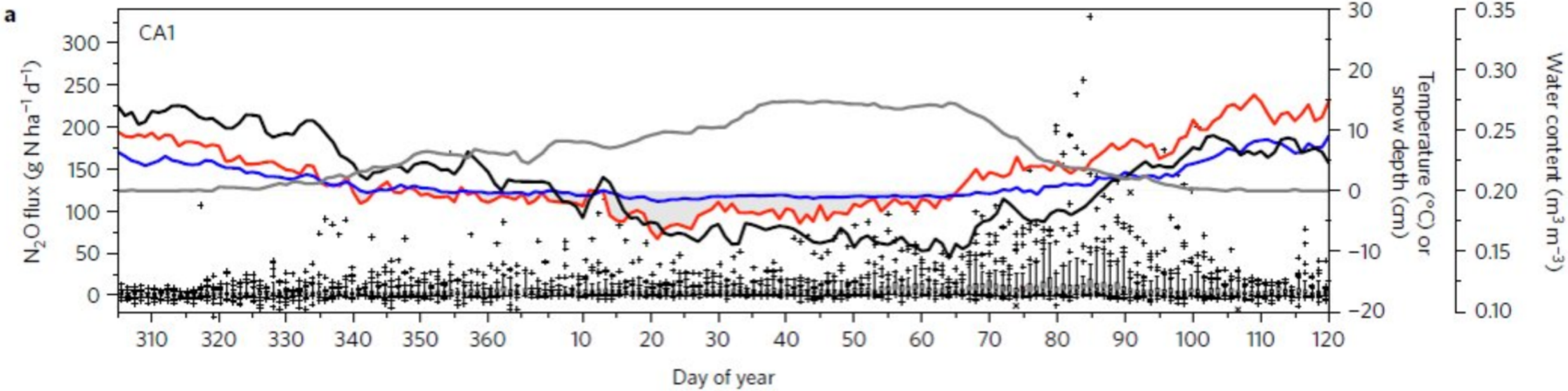
WHY NOT?



Data sources: National Spelling Bee and Centers for Disease Control & Prevention

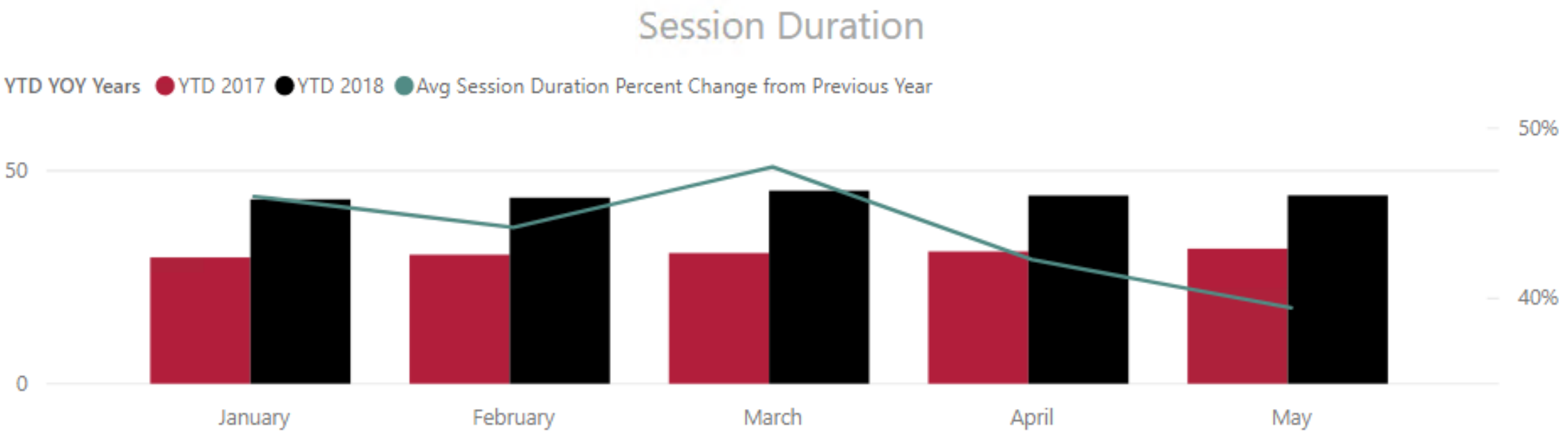

0.10
0.08
0.06
0.04
0.02
0.00
Conception Risk
6
5
4
3
2
1
Political Conservatism
Conception Risk
Political Conservatism
2
4
6
8
10
12
14
16
18
20
22
24
26
28
Menstrual Cycle Day

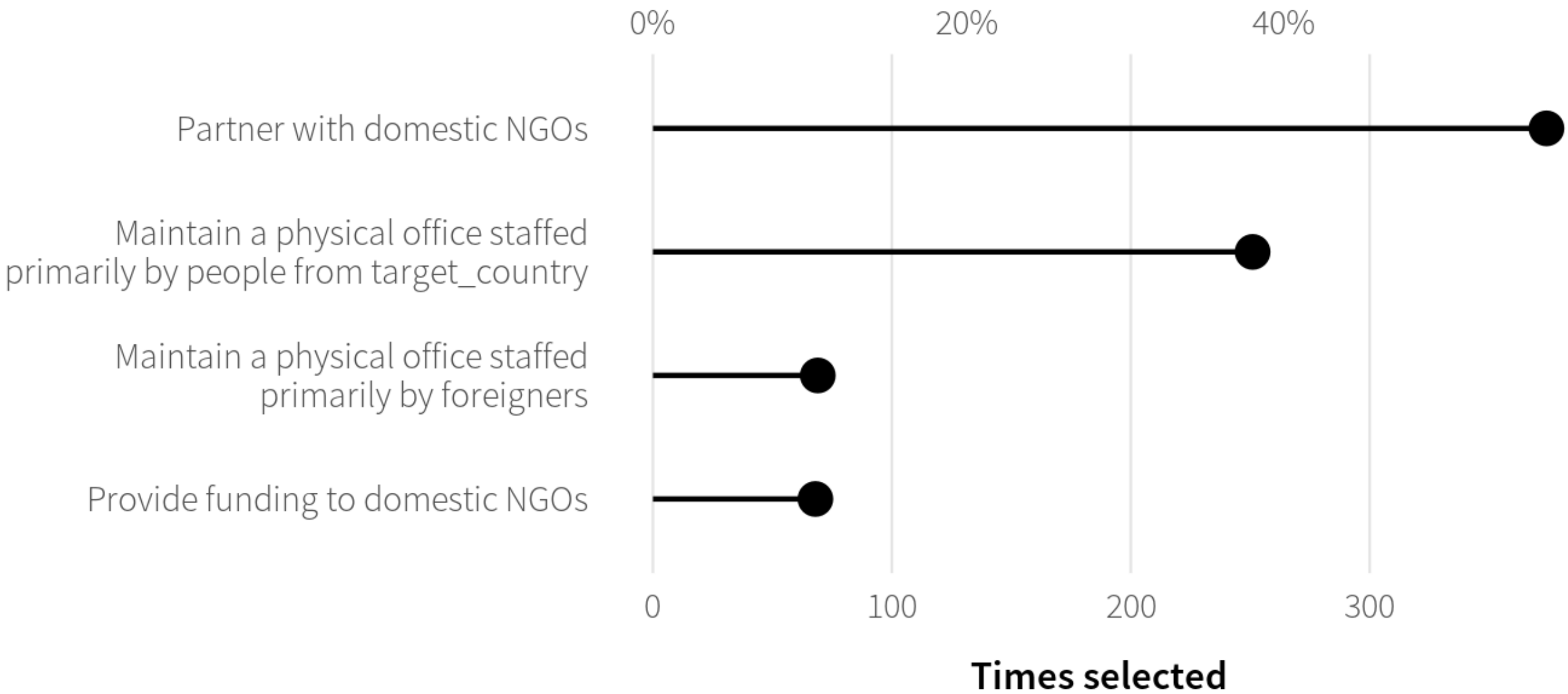

Labor share
.56
.58
.6
.62
.64
.66
Employment ratio
.8
.85
.9
.95
Output per worker growth
0
1
2
3
4
1945
1955
1965
1975
1985
1995
2005
2015
Year
Employment to population ratio for man, left axis
Labor share in nonfarm business, left axis
Labor productivity growth in nonfarm business, right axis

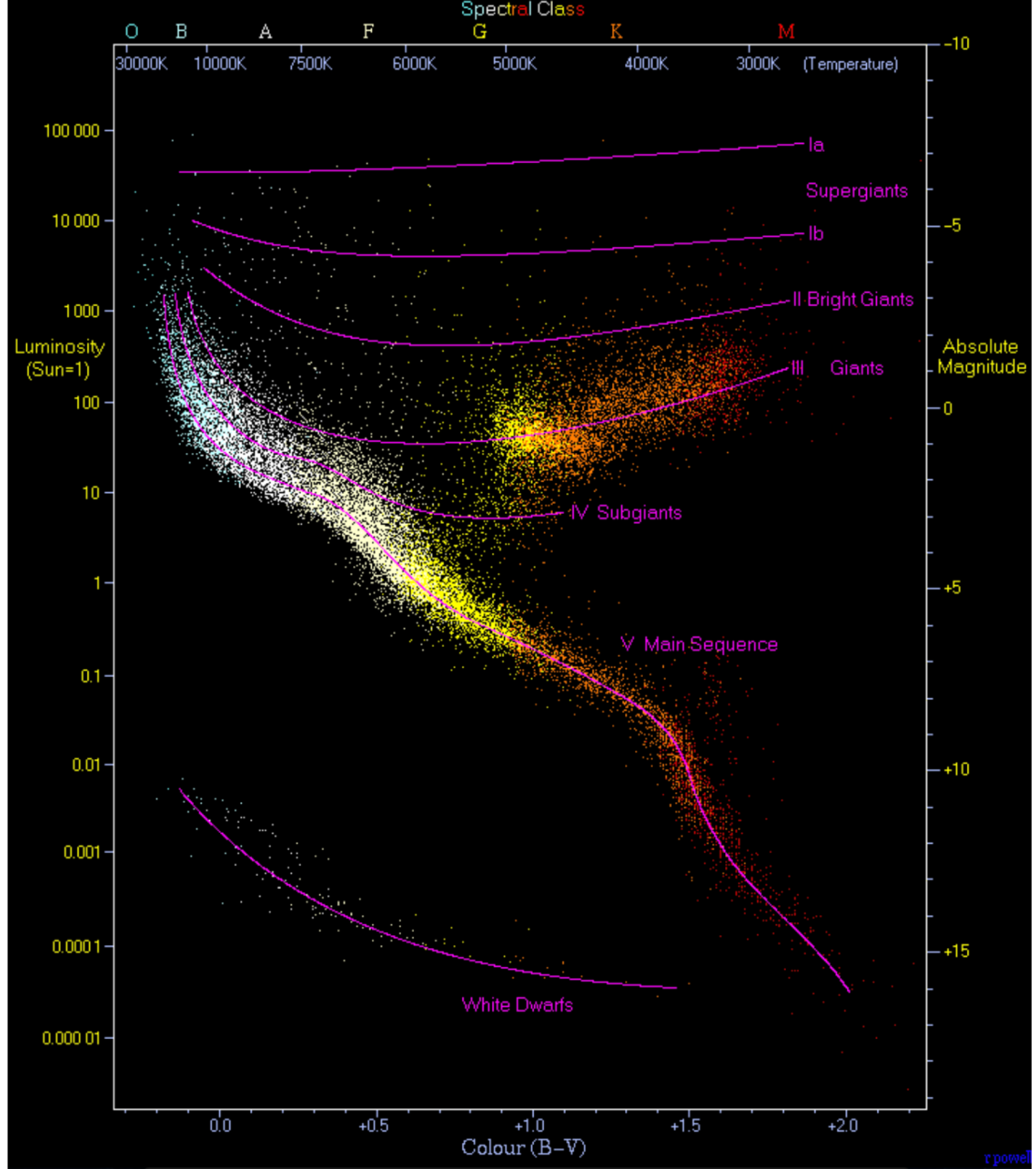

a
CA1
N_2O flux (g N ha^{-1} d^{-1})
300
250
200
150
100
50
0
30
20
10
0
−10
−20
Temperature (°C) or snow depth (cm)
0.35
0.30
0.25
0.20
0.15
0.10
Water content (m^3 m^{-3})
310
320
330
340
350
360
10
20
30
40
50
60
70
80
90
100
110
120
Day of year

WHEN IS IT LEGAL?

When the two y-axes measure the same thing

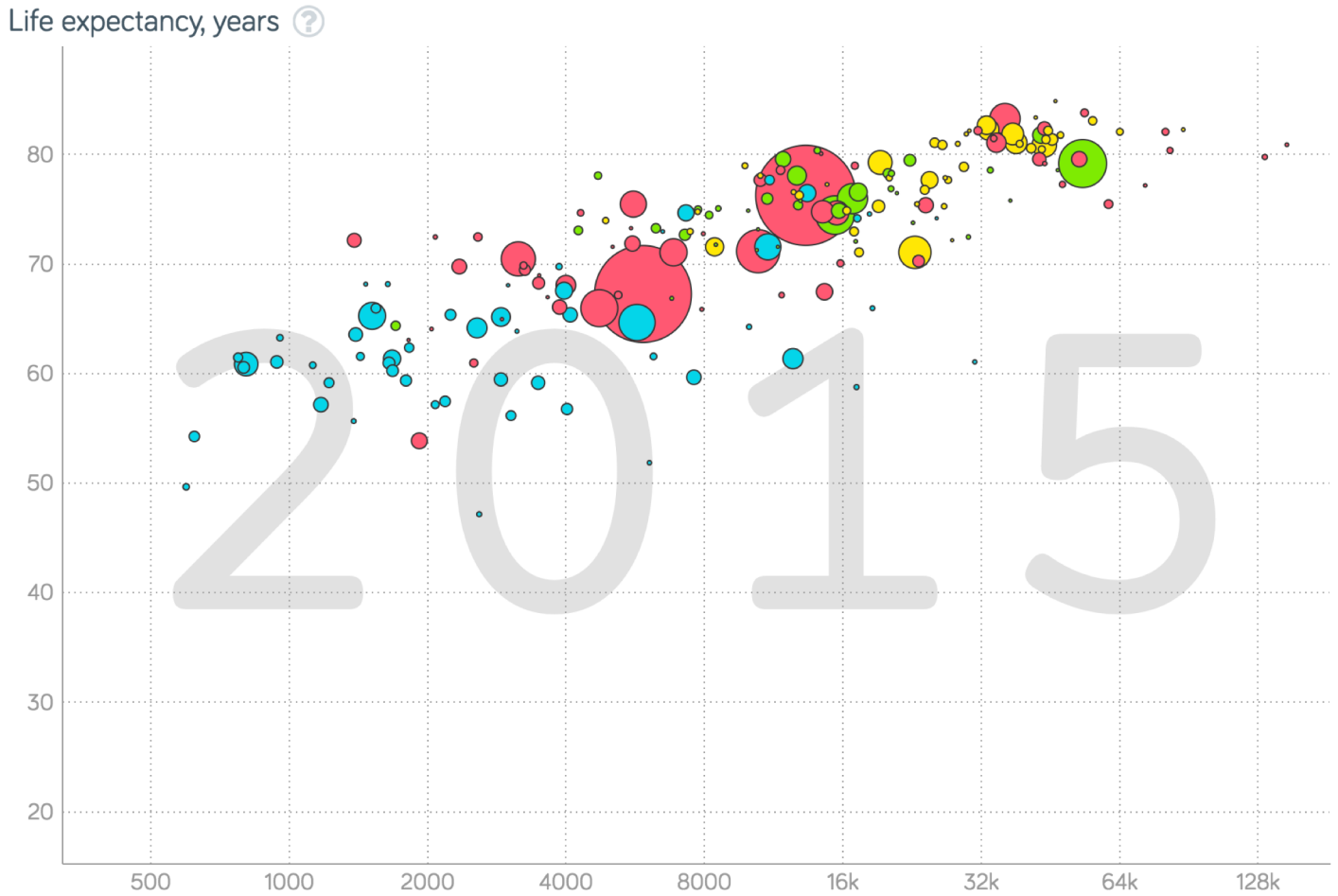



0%
20%
40%
Partner with domestic NGOs
Maintain a physical office staffed primarily by people from target_country
Maintain a physical office staffed primarily by foreigners
Provide funding to domestic NGOs
0
100
200
300
Times selected

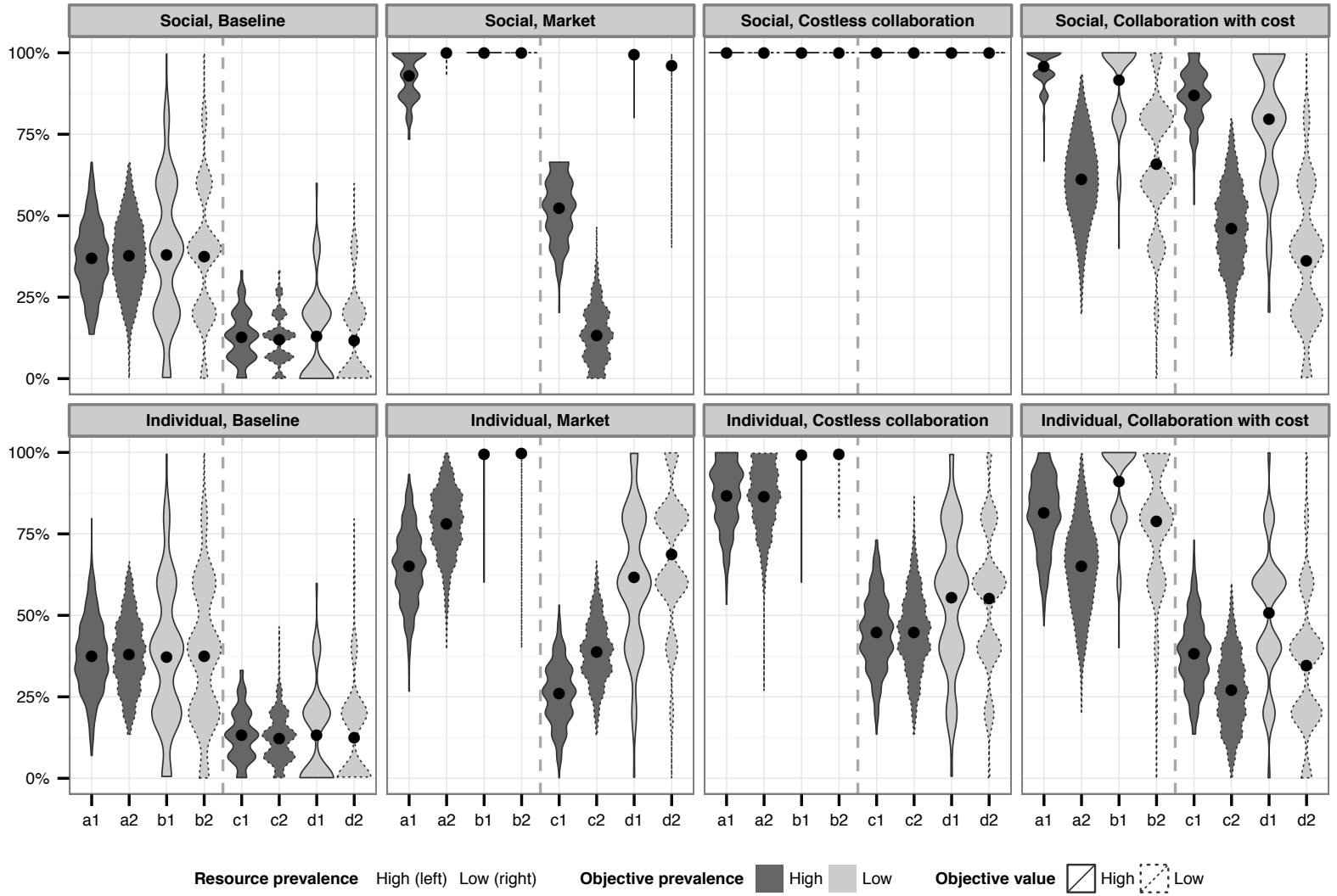

Spectral Class
O
B
A
F
G
K
M
30000K
10000K
7500K
6000K
5000K
4000K
3000K
(Temperature)
100 000
10 000
1000
100
10
1
0.1
0.01
0.001
0.0001
0.000 01
Luminosity
(Sun=1)
-10
-5
0
+5
+10
+15
Absolute
Magnitude
Ia
Supergiants
Ib
II Bright Giants
III Giants
IV Subgiants
V Main Sequence
White Dwarfs
0.0
+0.5
+1.0
+1.5
+2.0
Colour (B–V)
rpowell

ALTERNATIVES

Map variable to another aesthetic

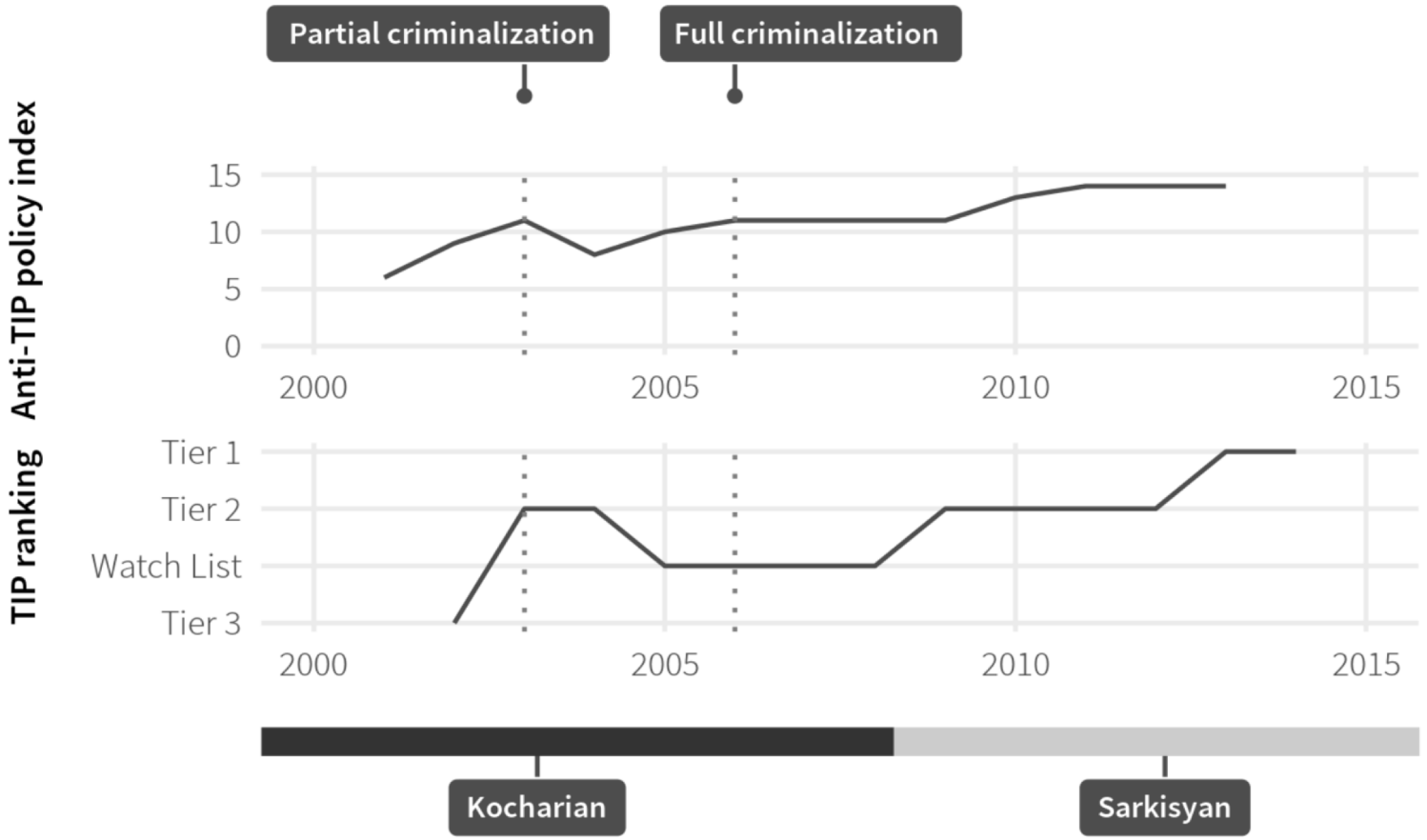


(NOT TOO MUCH THOUGH)



ALTERNATIVES

Use multiple plots



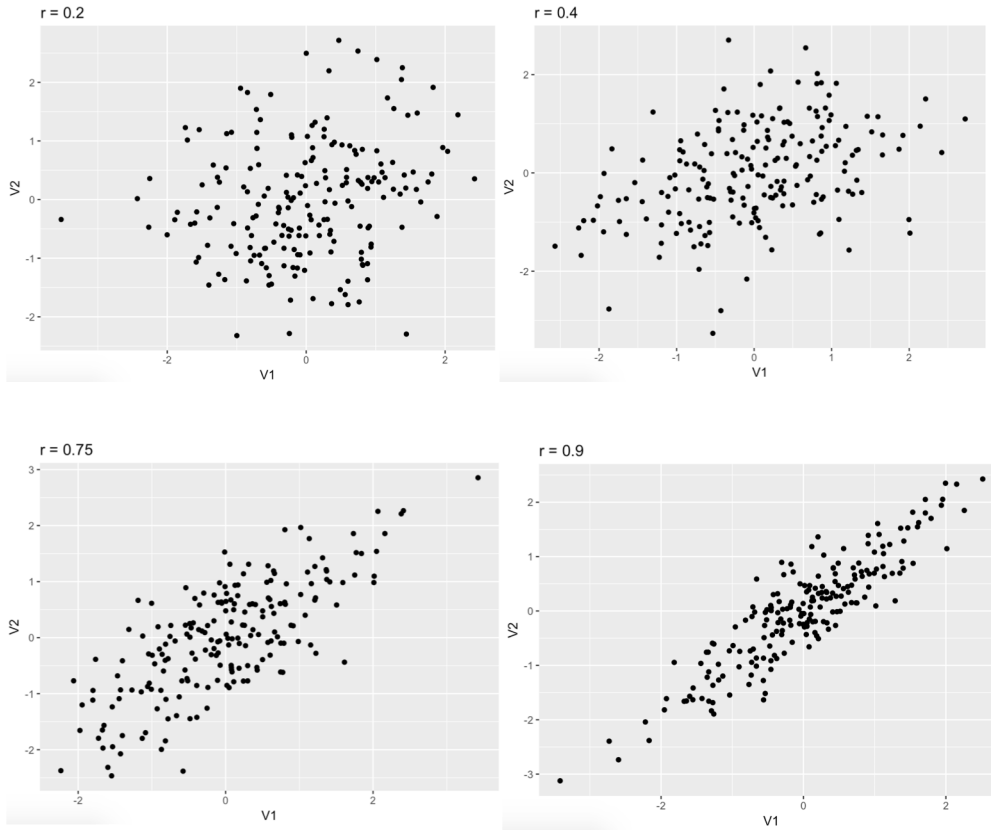
CORRELATION AND REGRESSION

WHAT IS *r*?

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

±0.1–0.3	Modest
±0.3–0.5	Moderate
±0.5–0.8	Strong
±0.8–0.9	Very strong

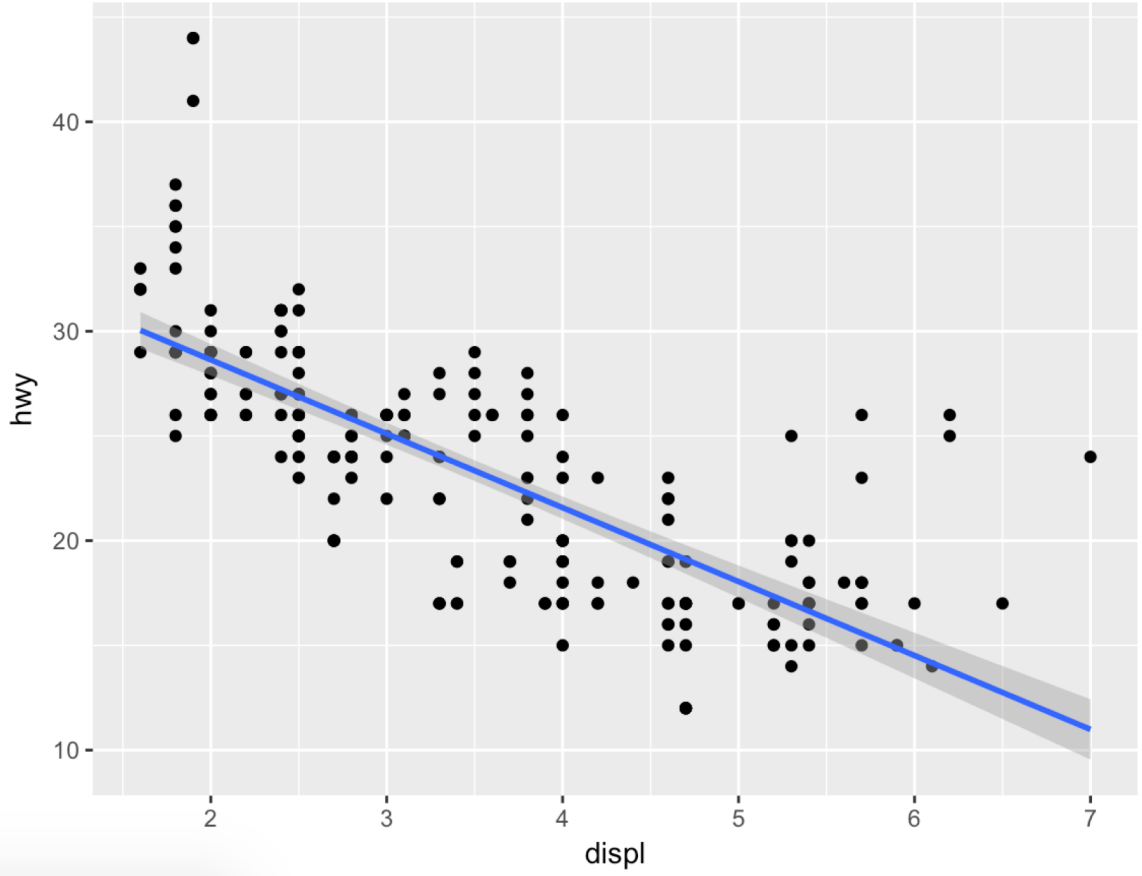


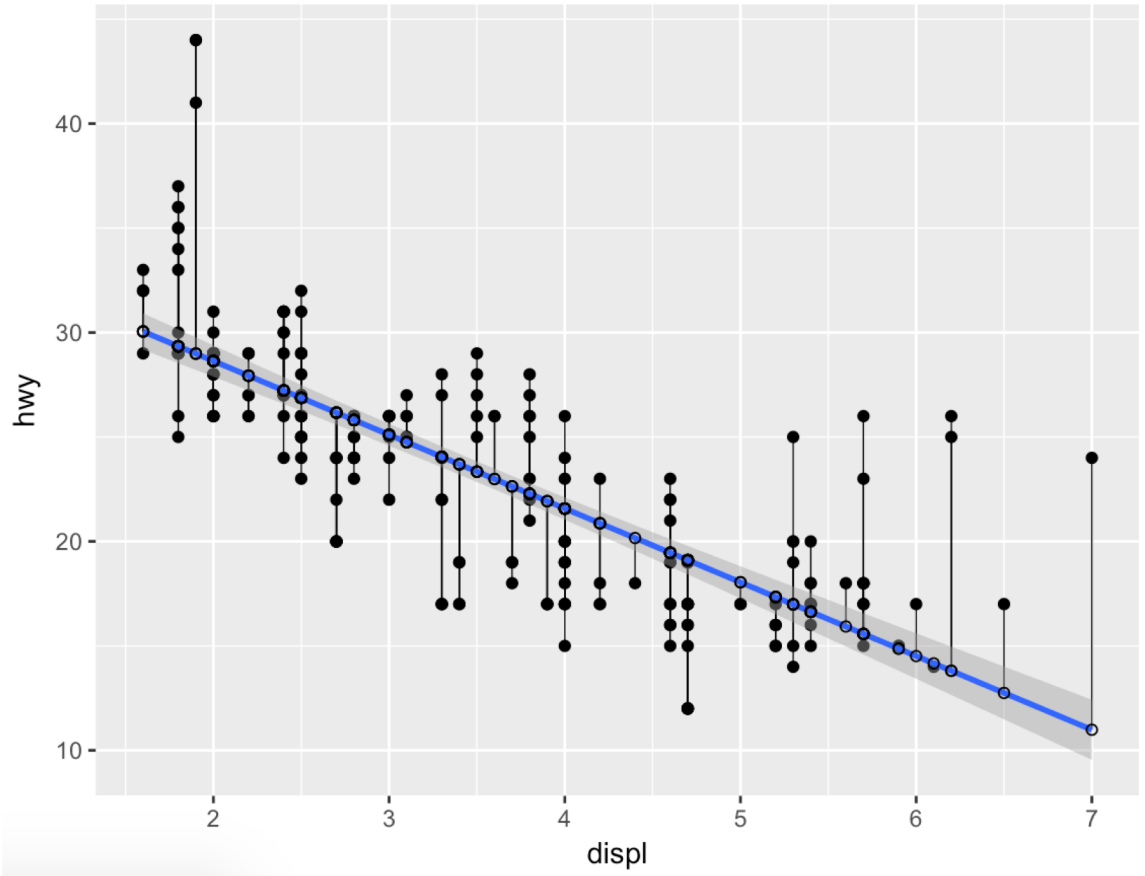
HOW TO INTERPRET

As the value of X goes up, Y tends to go up (or down) a lot/a little/not at all

Says nothing about how much Y changes when X changes

REGRESSION





REGRESSION

$$y = mx + b \qquad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

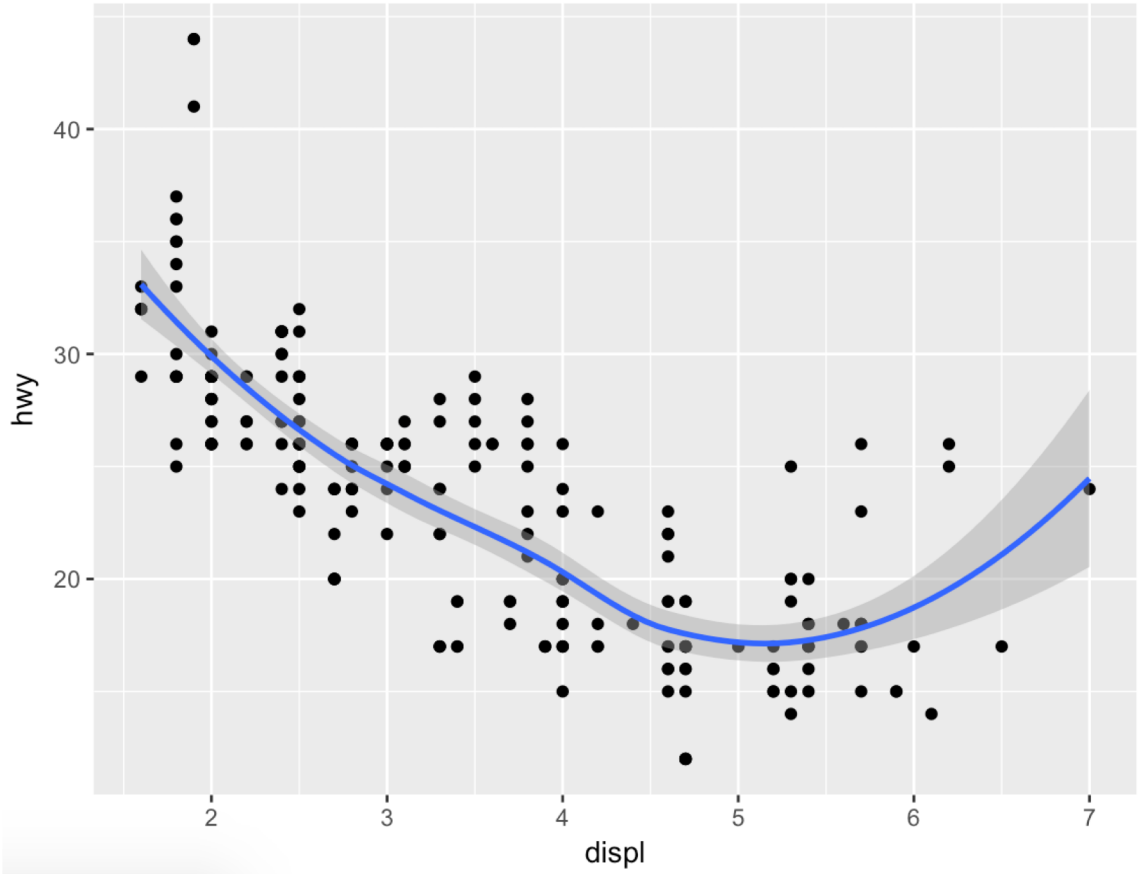
```
Call:
lm(formula = hwy ~ displ, data = mpg)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1039 -2.1646 -0.2242  2.0589 15.0105

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.6977     0.7204   49.55   <2e-16 ***
displ        -3.5306     0.1945  -18.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.836 on 232 degrees of freedom
Multiple R-squared:  0.5868,    Adjusted R-squared:  0.585
F-statistic: 329.5 on 1 and 232 DF,  p-value: < 2.2e-16
```

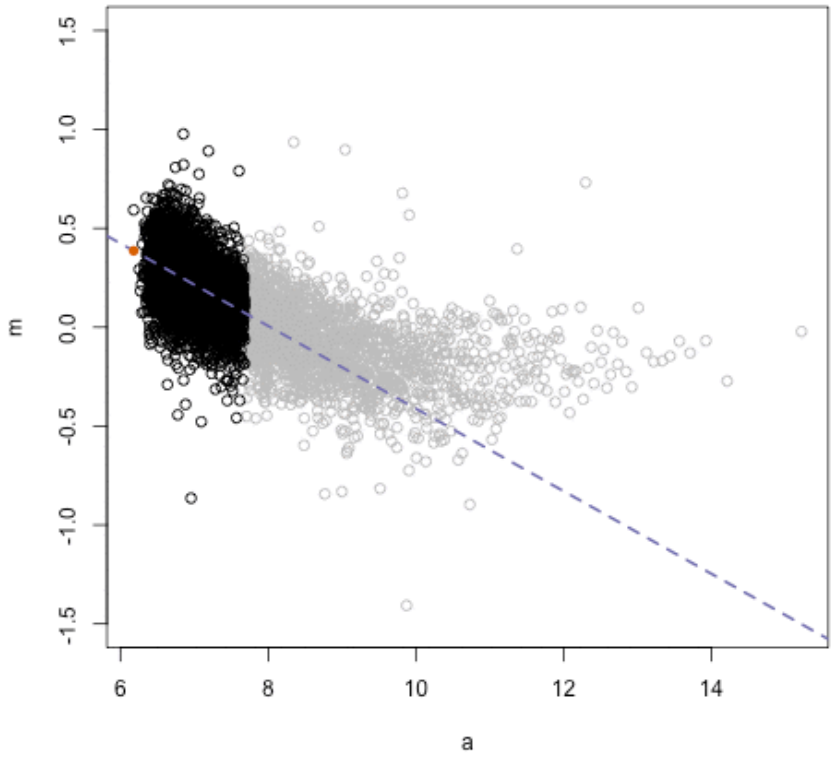
```
Call:
lm(formula = hwy ~ displ + cyl + drv, data = mpg)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7095 -2.0282 -0.1297  1.3760 13.8110

Coefficients:
            Estimate Std. Error t value       Pr(>|t|)
(Intercept)  33.0915     1.0306  32.108        < 2e-16 ***
displ        -1.1245     0.4614  -2.437         0.0156 *
cyl          -1.4526     0.3334  -4.357 0.000019922184 ***
drvf          5.0446     0.5134   9.826        < 2e-16 ***
drvr          4.8851     0.7116   6.864 0.000000000062 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.968 on 229 degrees of freedom
Multiple R-squared:  0.7559,    Adjusted R-squared:  0.7516
F-statistic: 177.2 on 4 and 229 DF,  p-value: < 2.2e-16
```

HOW TO INTERPRET

As X goes up by one unit, there's an associated change in Y of β_1

OTHER LINES





Problems: **Math is hard** **Overfitting**

CORRELATION AND CAUSATION

GODWIN'S LAW FOR STATISTICS

Correlation does not imply causation

Except when it does

Even if it doesn't,
this phrase is useless
and kills discussion

Not everyone found the news believable. “Facepalm. Correlation doesn’t imply causation,” wrote one unhappy Internet user. “That’s pretty much how I read this too… correlation is NOT causation,” agreed a Huffington Post superuser, seemingly distraught. “I was surprised not to find a discussion of correlation vs. causation,” cried someone at Hacker News. “Correlation does not mean causation,” a reader moaned at Slashdot. “There are so many variables here that it isn’t funny.”

“It seems everybody these days is so smart about how correlation doesn't equal causation, but all they can do is blurt that out, leaving the table empty of solutions, ideas, and values”

Katie Olson

David Robinson
@drob

Following

Correlation implies causation, don't @ me

1:12 PM - 22 Jun 2017 from Manhattan, NY

4 Retweets **56** Likes

2 4 56

Tweet your reply

David Robinson @drob · 22 Jun 2017

Replying to @drob

"Correlation implies casuation," the dean whispered as he handed me my PhD.

"But then why-"

"Because if they knew, they wouldn't need us."

5 46 169

John B. Holbein @JohnHolbein1 · Apr 7

Causality isn't achieved; it's approached.

3 1 8

Show this thread

John B. Holbein @JohnHolbein1 · Apr 7

Causality isn't binary; it's a continuum.

1 5 13

Show this thread

CORRELATION VS. CAUSATION

How do we figure out correlation?

Math and statistics

How do we figure out causation?

Philosophy. No math.

THE CAUSALITY CONTINUUM

Differences

Pre-post

Multiple regression

Matching

Diff-in-diff

Natural experiments

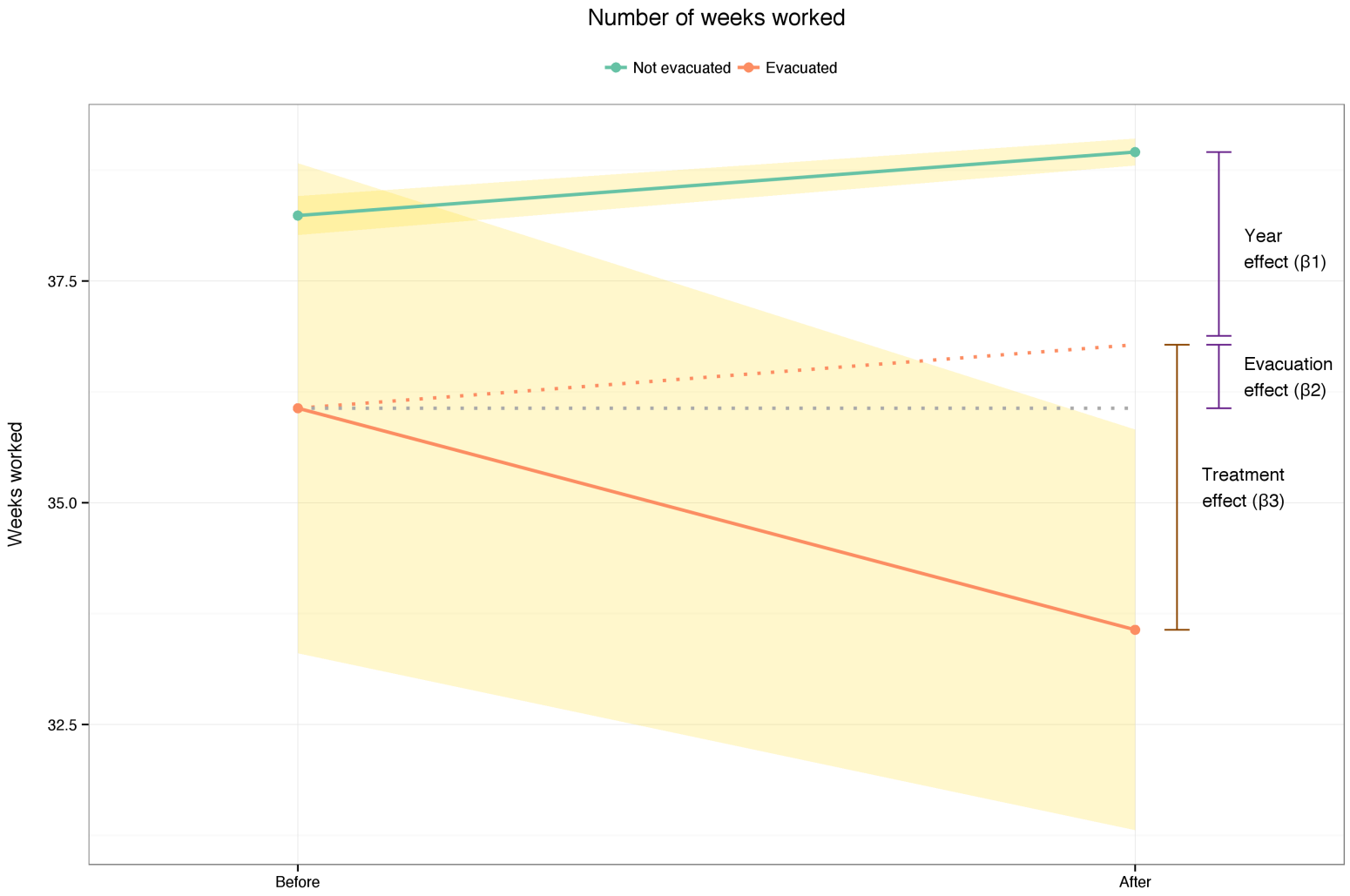
Regression discontinuity

RCTs

Correlation

Causation

DIFF-IN-DIFF



Effect of Katrina on employment

Weeks worked ~
Year +
Evacuated +
Year × Evacuated

MULTIPLE REGRESSION

Table 2: OLS models for four standardized tests

VARIABLES	(1) Reading	(2) Math	(3) Listening	(4) Words
Small class	6.47***	8.84***	3.24**	6.99***
	(1.45)	(2.32)	(1.42)	(1.60)
Regular + aide class	1.00	0.42	-0.58	1.27
	(1.26)	(2.14)	(1.32)	(1.42)
White or Asian	7.85***	16.91***	17.98***	7.08***
	(1.61)	(2.40)	(1.70)	(1.91)
Girl	5.39***	6.46***	2.67***	5.03***
	(0.78)	(1.12)	(0.74)	(0.94)
Free/reduced lunch	-14.69***	-20.08***	-15.23***	-15.97***
	(0.91)	(1.33)	(0.90)	(1.07)
Teacher white or Asian	-0.56	-1.01	-3.68	0.46
	(2.66)	(3.80)	(2.59)	(3.07)
Years of teacher experience	0.30**	0.42**	0.25*	0.30**
	(0.12)	(0.20)	(0.15)	(0.14)
Teacher has MA	-0.75	-2.20	0.50	0.24
	(1.25)	(2.08)	(1.24)	(1.46)
School fixed effects	X	X	X	X
Constant	431.69***	475.52***	531.28***	428.97***
	(3.12)	(4.49)	(2.84)	(3.59)
Observations	5,728	5,809	5,776	5,790
R-squared	0.08	0.07	0.09	0.06
Number of schools	79	79	79	79

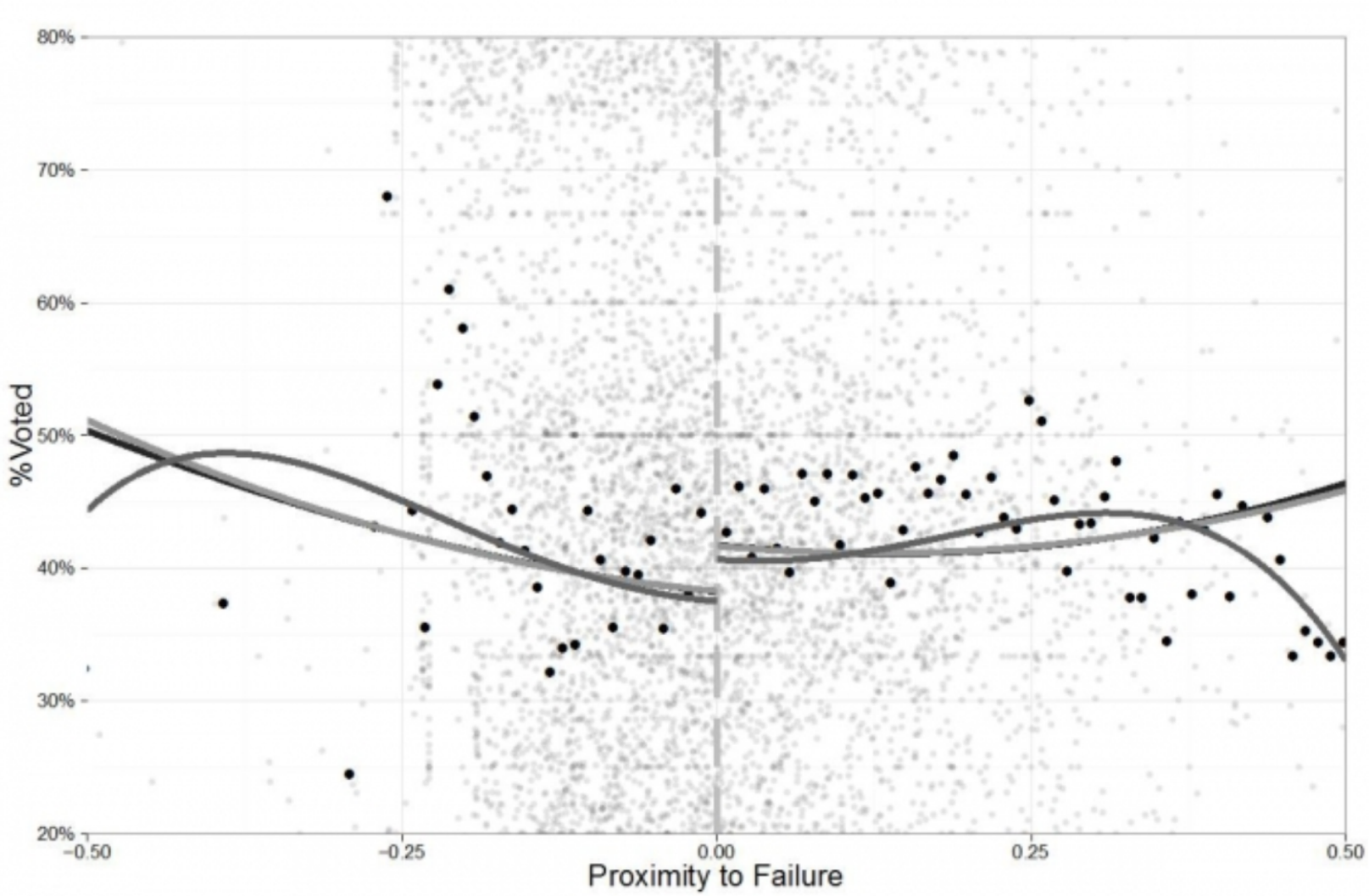
Robust standard errors in parentheses.
Standard errors corrected with Huber-White clustering by kindergarten teacher ID
*** p<0.01, ** p<0.05, * p<0.1

Warning!

Don't say that smaller class sizes cause increases in test scores

Do say "Smaller class sizes predict an increase in test scores, controlling for X, Y, and Z"

REGRESSION DISCONTINUITY



NATURAL EXPERIMENTS

Figure 3A: Additional Midnights: Before Law Change

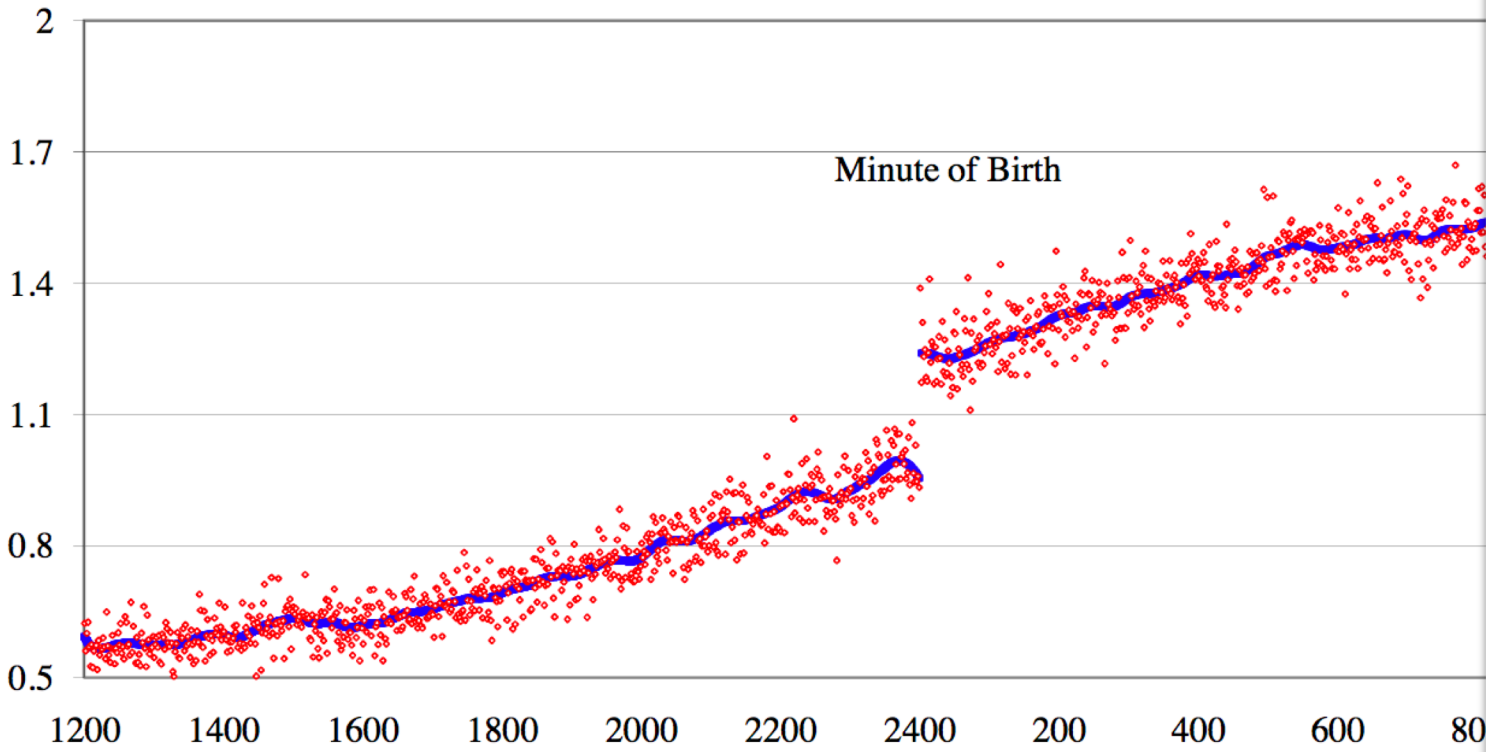


Figure 4A: 28-Day Readmission Rate: Before Law Change

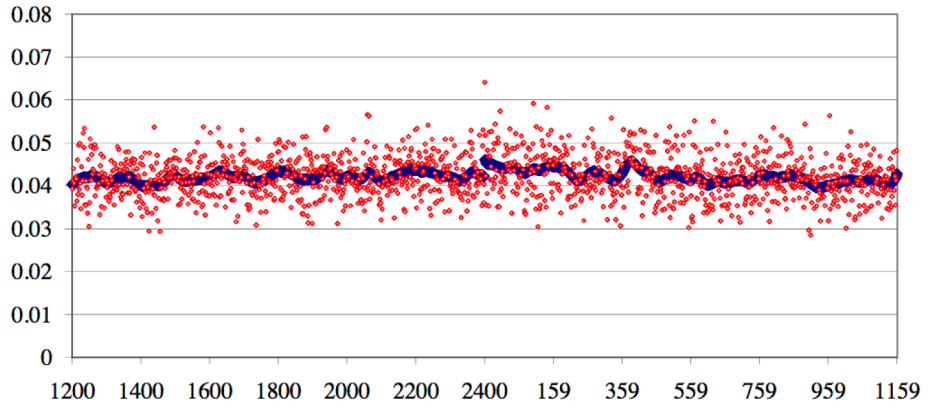
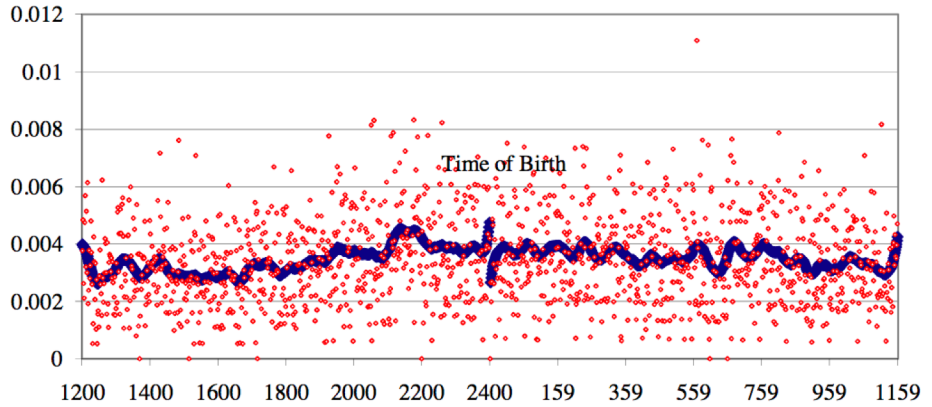
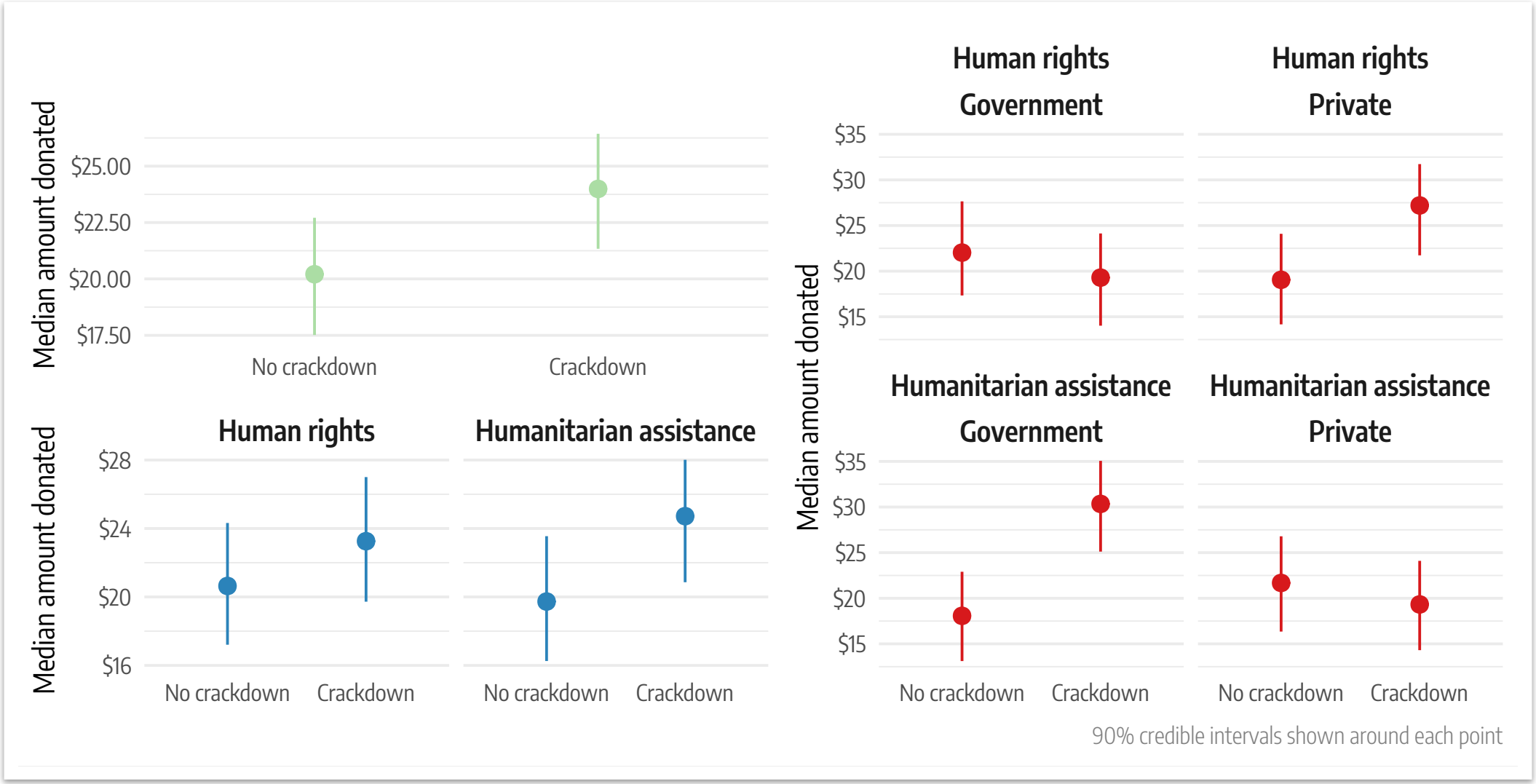


Figure 4C: 28-Day Mortality Rate: Before Law Change



LAB EXPERIMENTS



THE CAUSALITY CONTINUUM

Differences

Pre-post

Multiple regression

Matching

Diff-in-diff

Natural experiments

Regression discontinuity

RCTs

Correlation

Causation

GGPLOT EXAMPLES