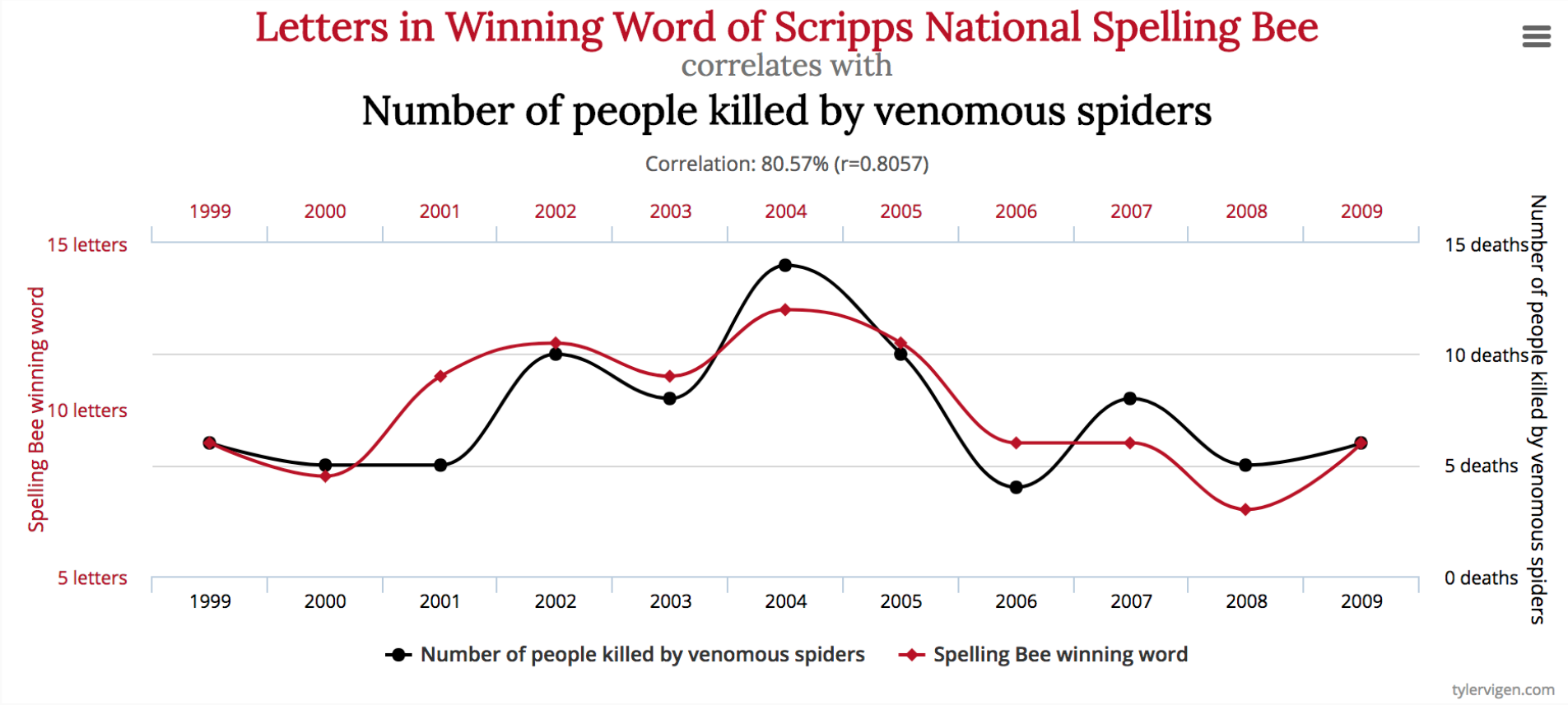
Relationships



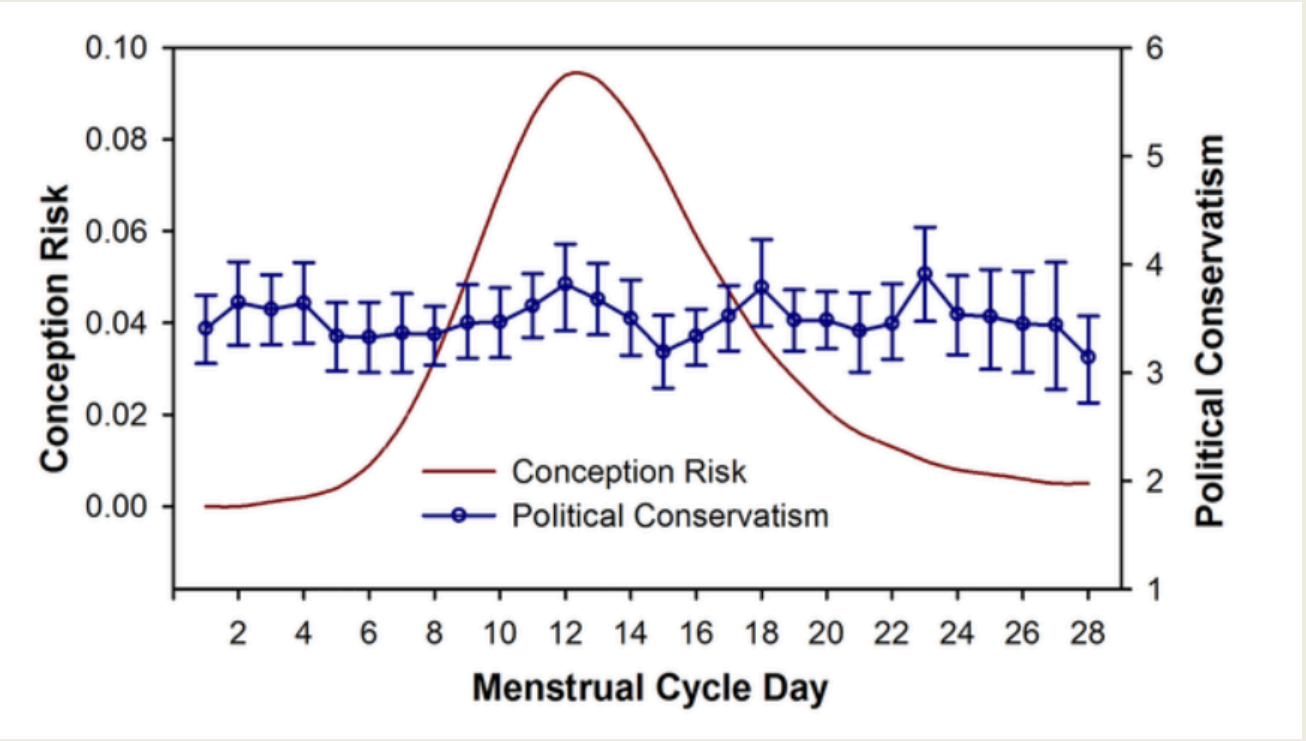
Data sources: National Spelling Bee and Centers for Disease Control & Prevention

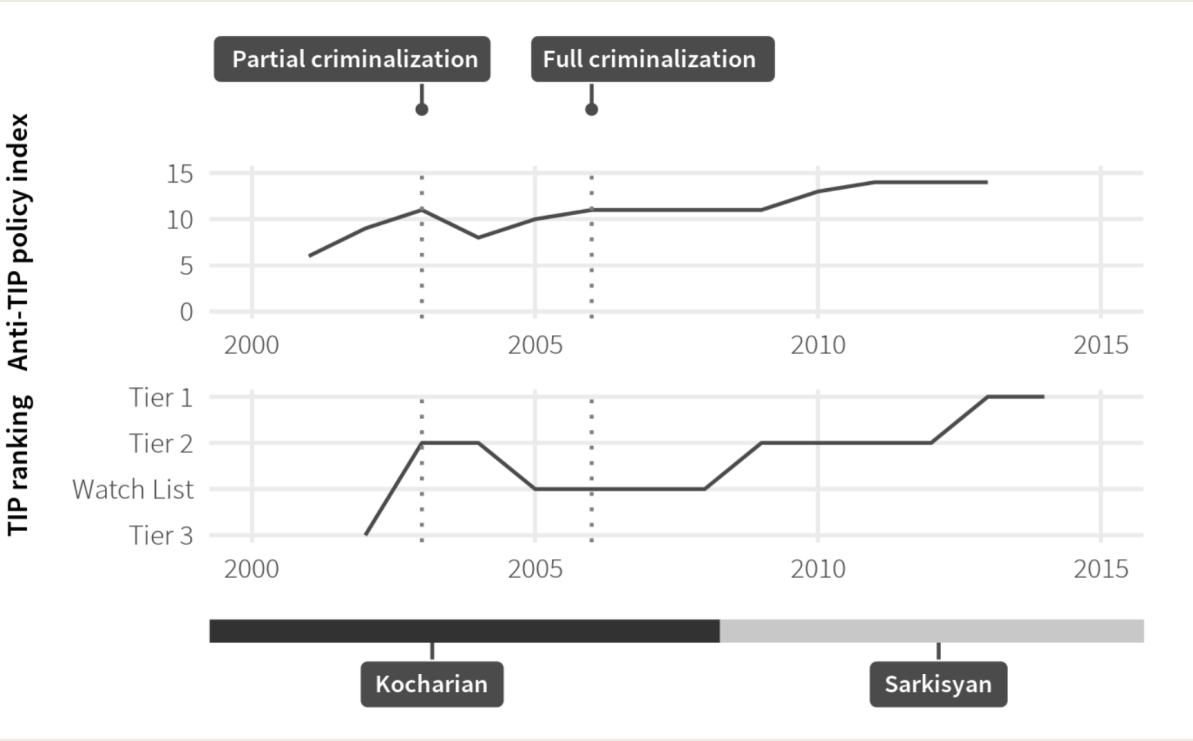
Relationships

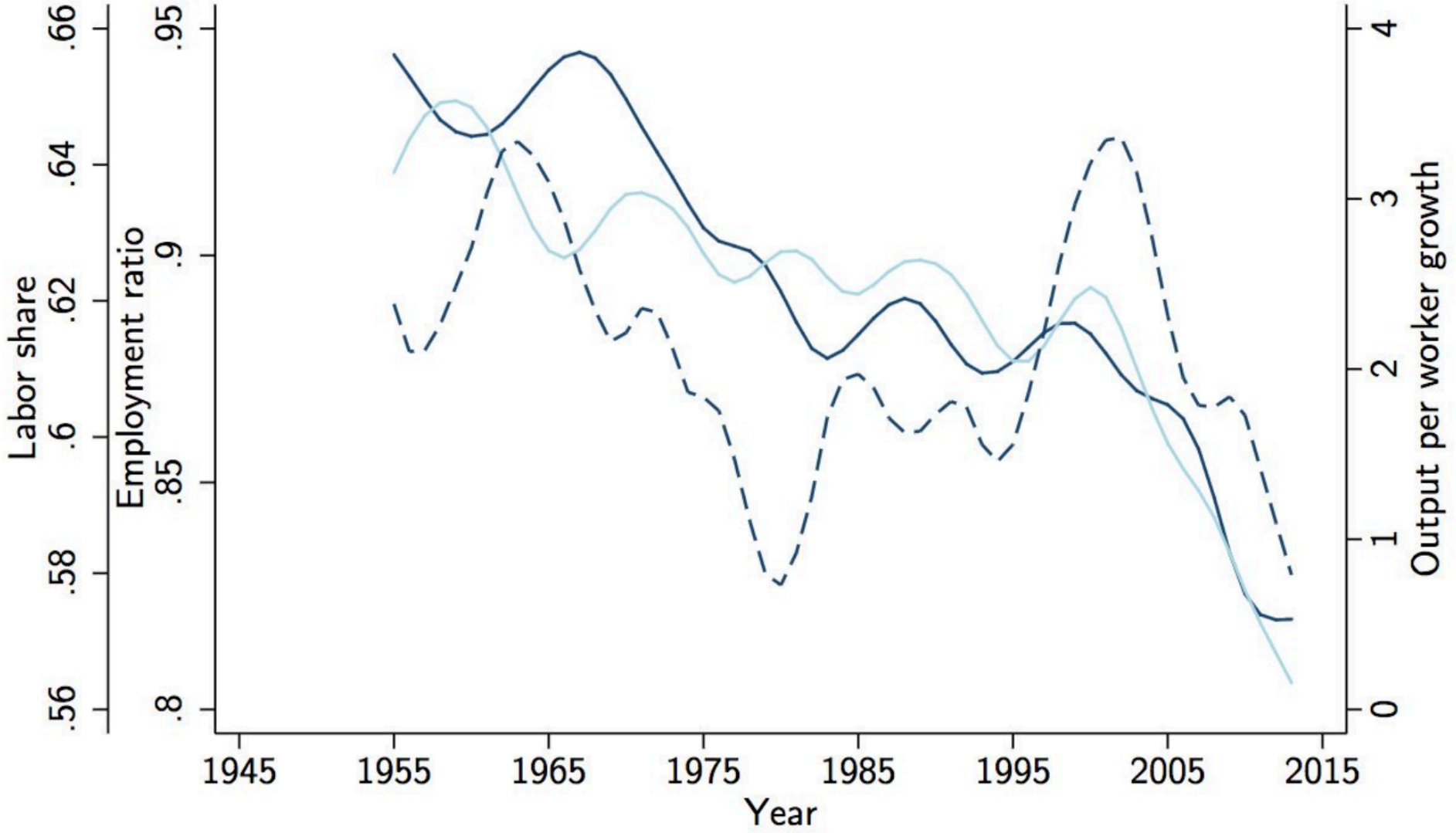
Stats 201

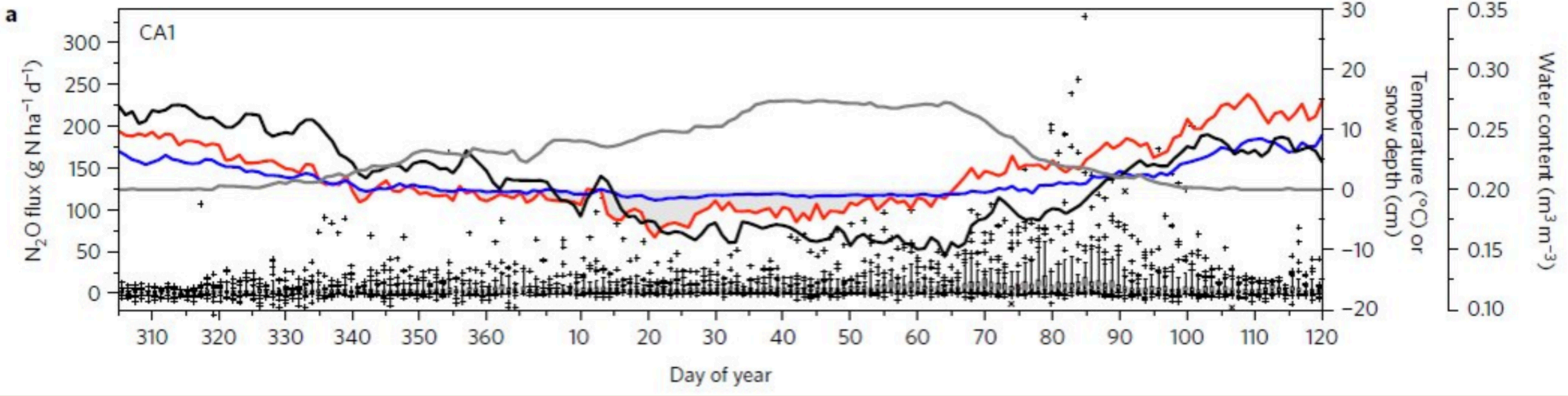
AdventuRe time

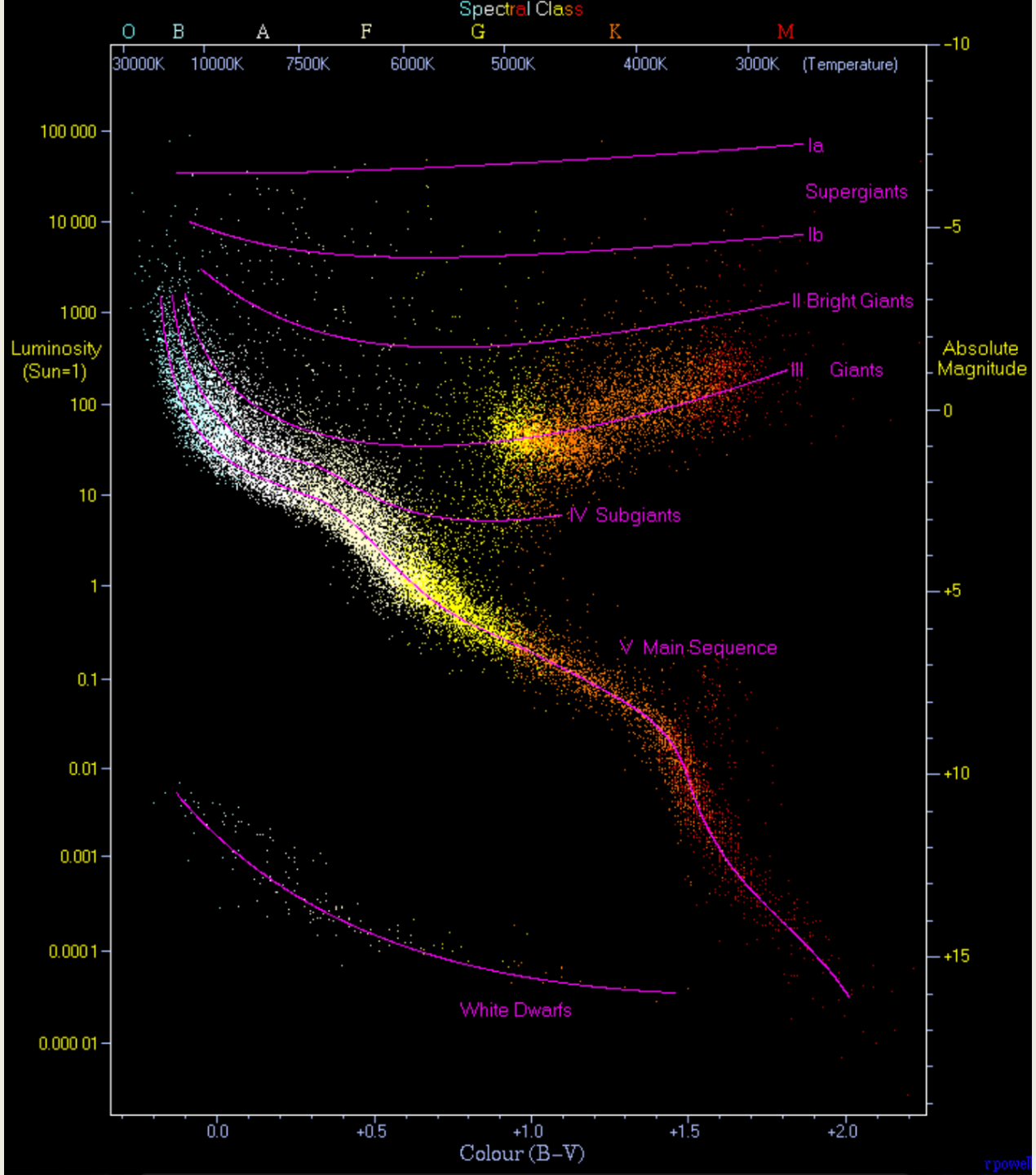
y-axes for everyone!

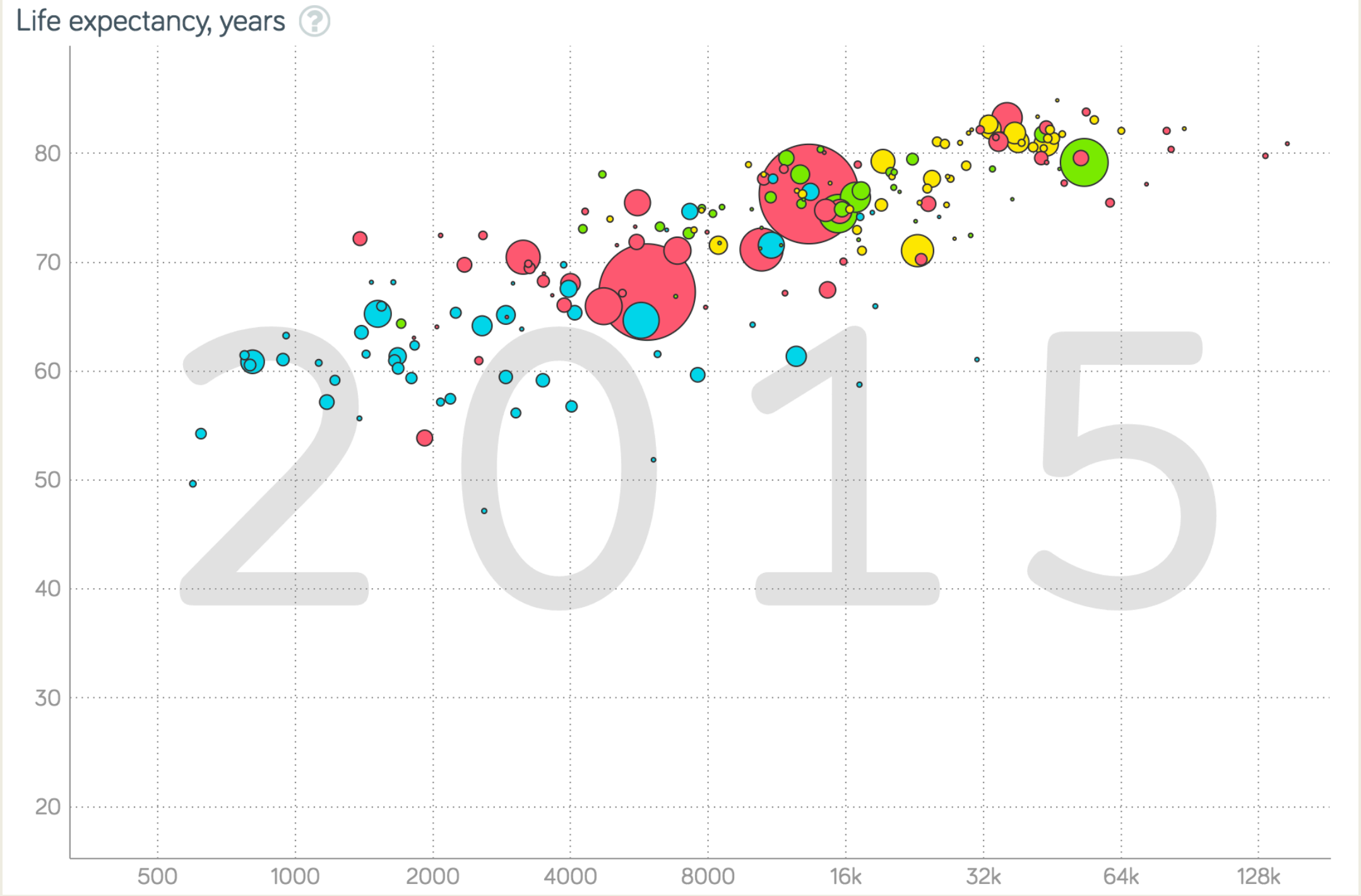





Labor share
.56
.58
.6
.62
.64
.66
Employment ratio
.8
.85
.9
.95
Output per worker growth
0
1
2
3
4
1945
1955
1965
1975
1985
1995
2005
2015
Year
Employment to population ratio for man, left axis
Labor share in nonfarm business, left axis
Labor productivity growth in nonfarm business, right axis


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


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

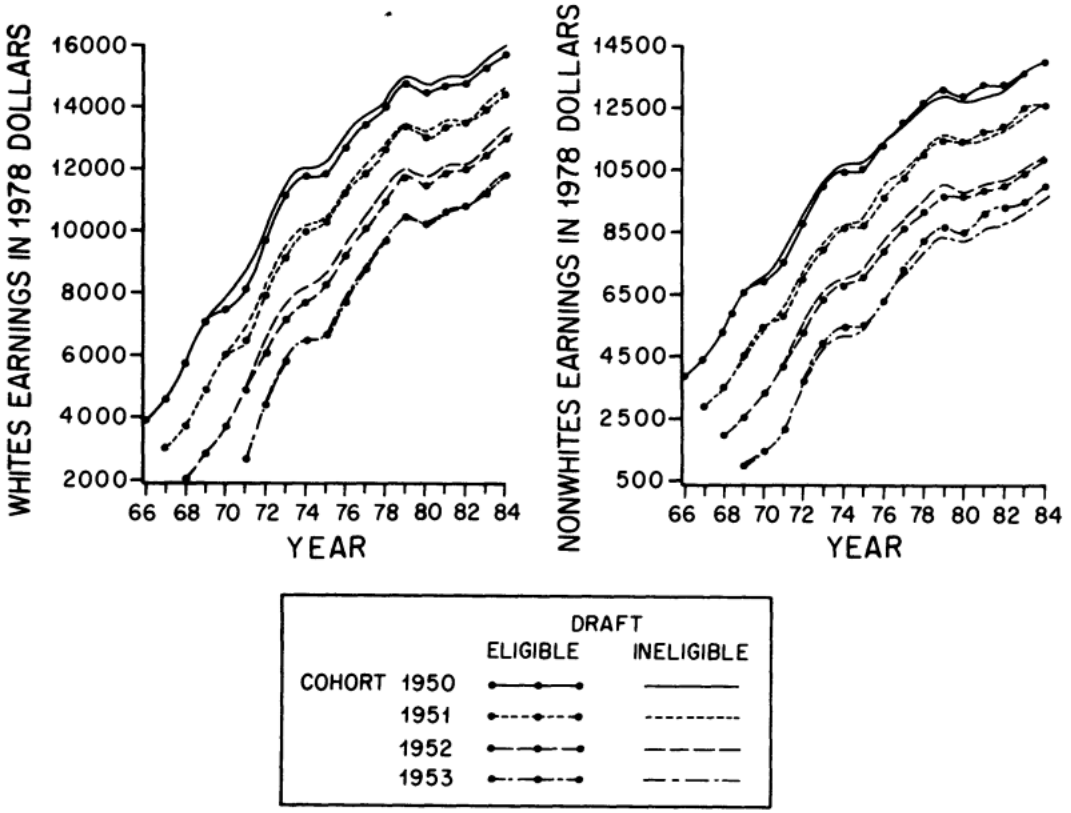

Life expectancy, years
80
70
60
50
40
30
20
2015
500
1000
2000
4000
8000
16k
32k
64k
128k

Obligatory correlation does not imply causation slide

But sometimes it does.

And even if it doesn't, it's not a helpful argument.

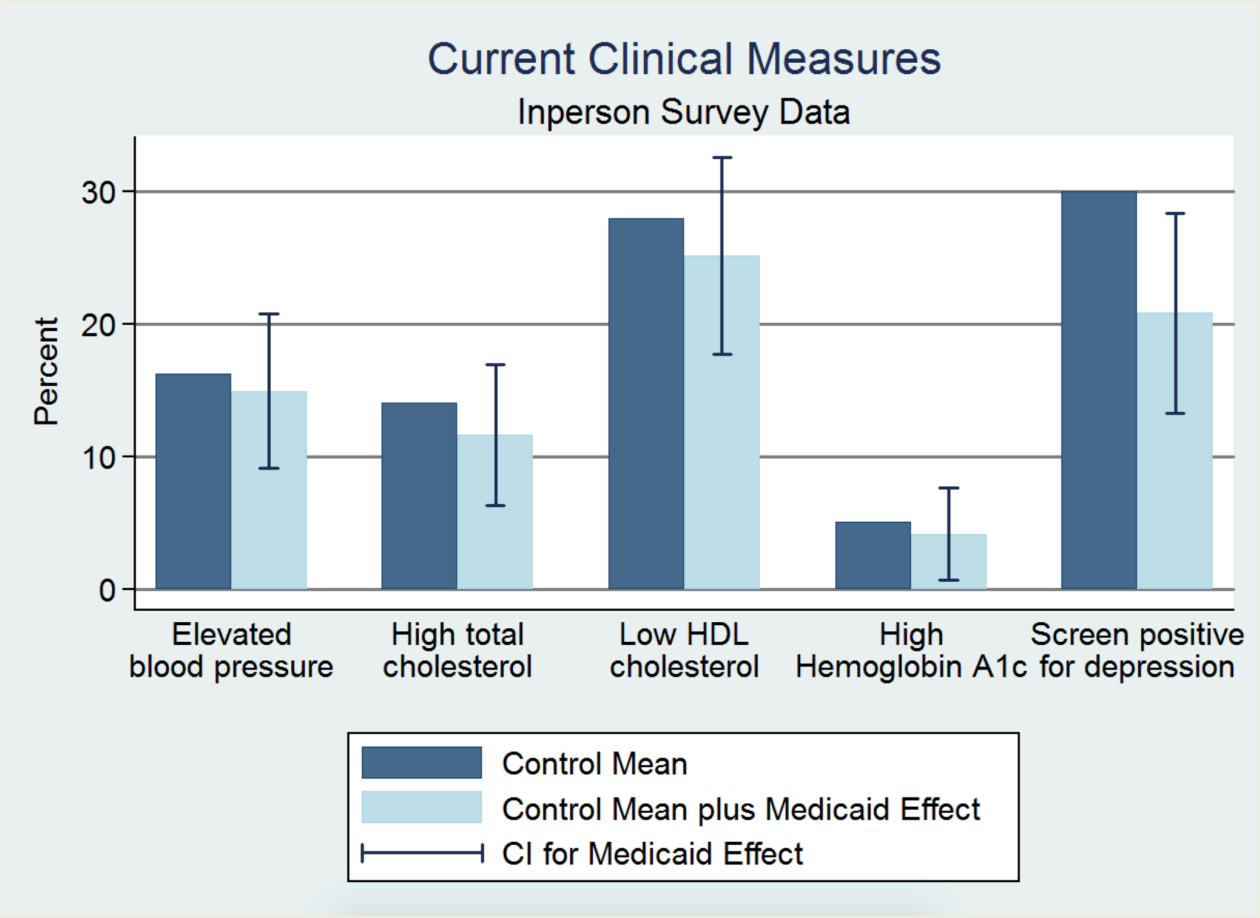
Not everyone found the news believable. "Facepalm. Correlation doesn't imply causation," wrote one unhappy Internet user. "That's pretty much how I read this too... correlation is NOT causation," agreed a Huffington Post superuser, seemingly distraught. "I was surprised not to find a discussion of correlation 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."

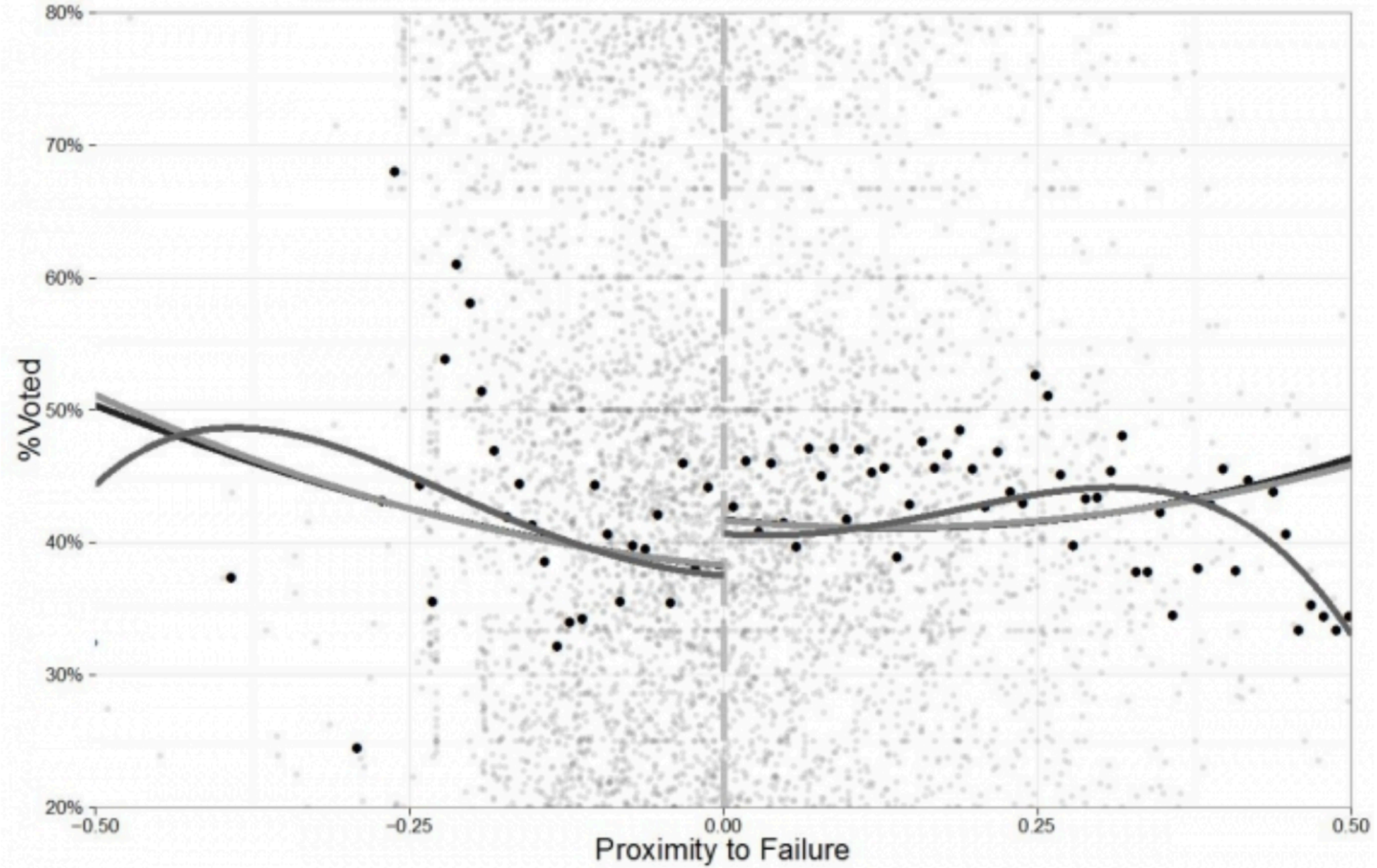


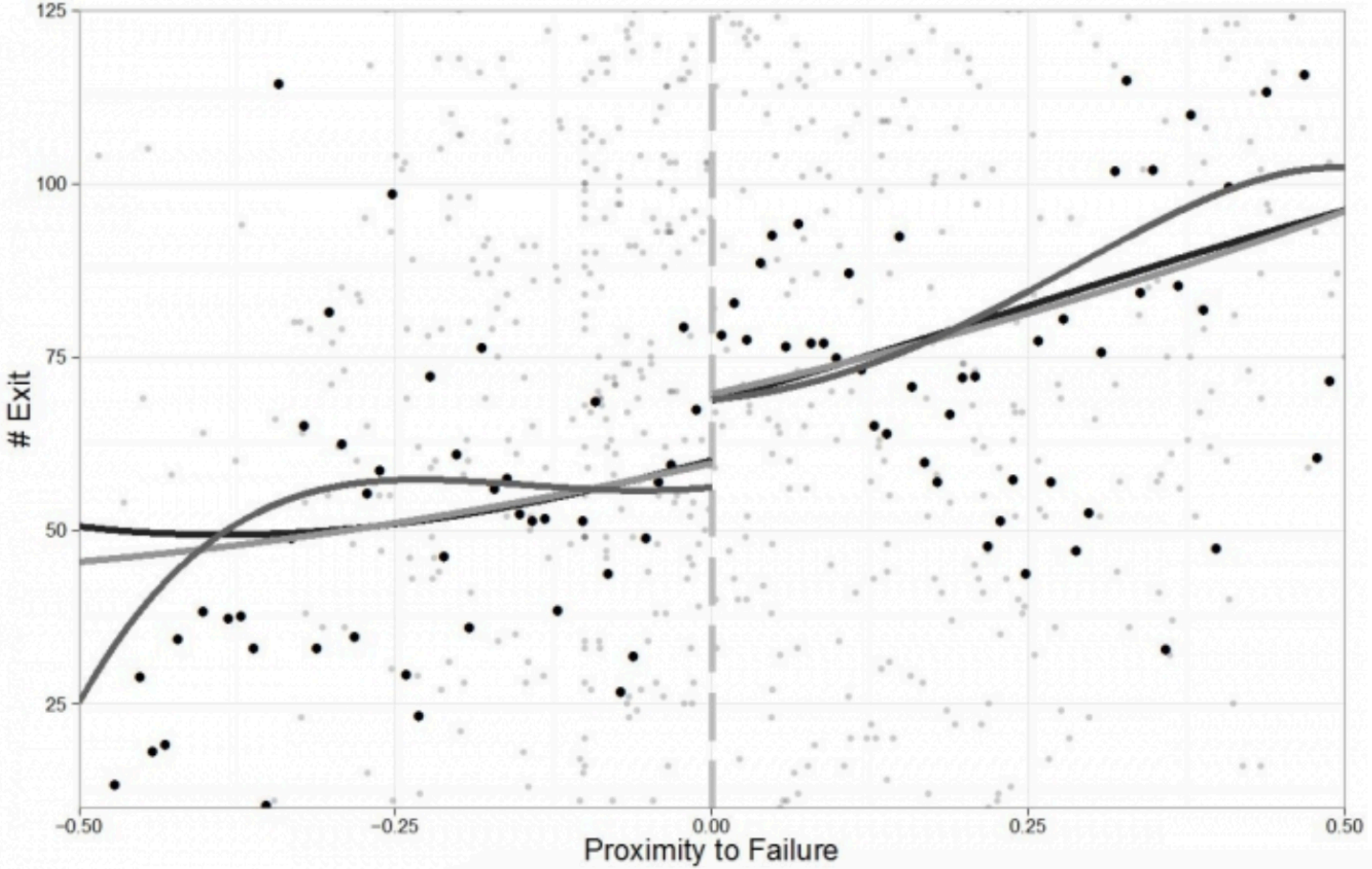
Notes: The figure plots mean W-2 compensation in 1981–84 against probabilities of veteran status by cohort and groups of five consecutive lottery numbers for white men born 1950–53. Plotted points consist of the average residuals (over four years of earnings) from regressions on period and cohort effects. The slope of the least squares regression line drawn through the points is −2,384 with a standard error of 778, and is an estimate of α in the equation

$$\bar{y}_{ctj} = \beta_c + \delta_t + \hat{p}_{cj}\alpha + \bar{u}_{ctj}.$$

FIGURE 3. EARNINGS AND THE PROBABILITY OF VETERAN STATUS BY LOTTERY NUMBER




80%
70%
60%
50%
40%
30%
20%
% Voted
−0.50
−0.25
0.00
0.25
0.50
Proximity to Failure


125
100
75
50
25
Exit
−0.50
−0.25
0.00
0.25
0.50
Proximity to Failure

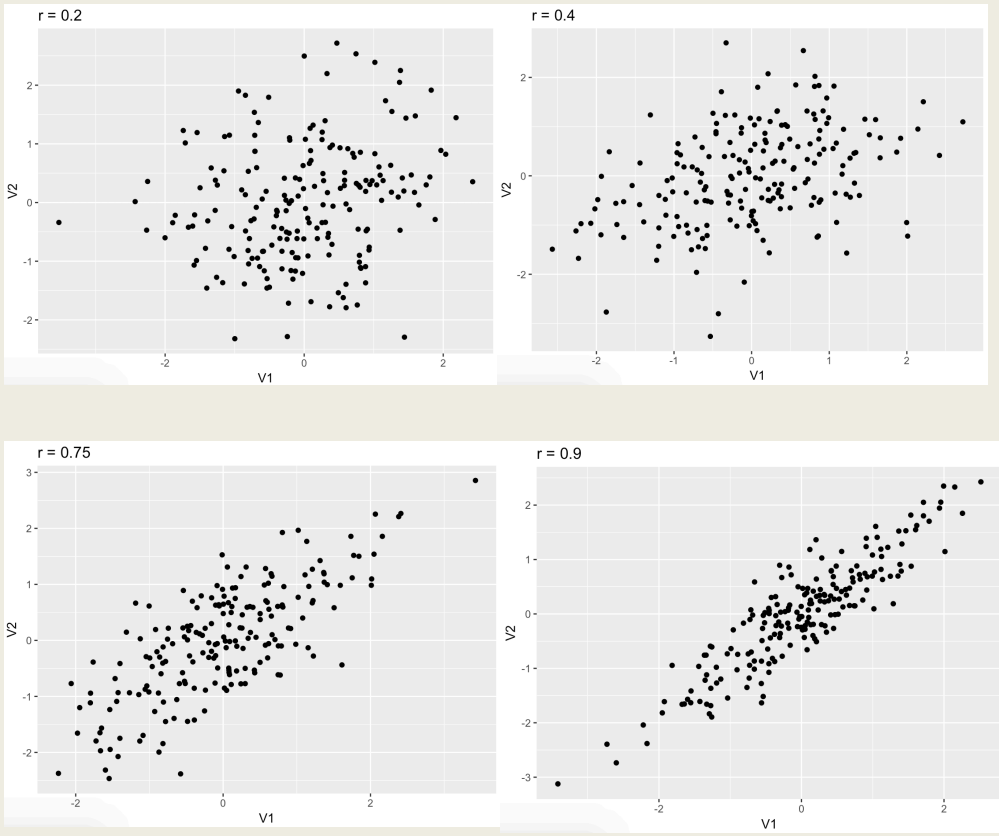
⚡Stats 201⚡

What is *r*?

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

What does *r* mean?

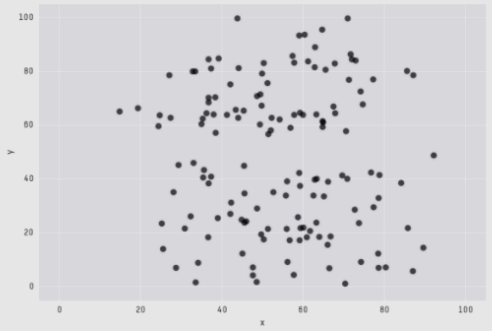
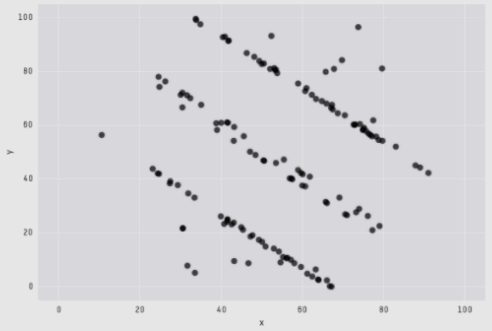
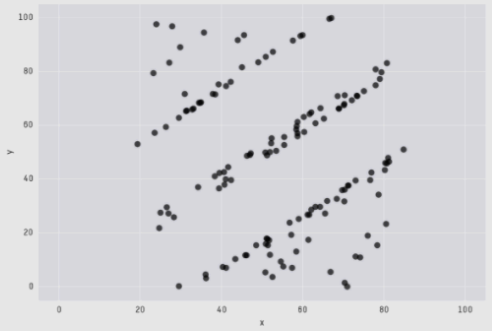
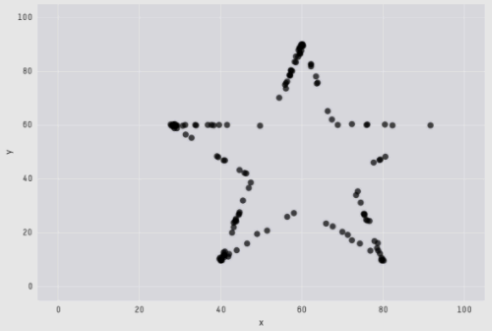
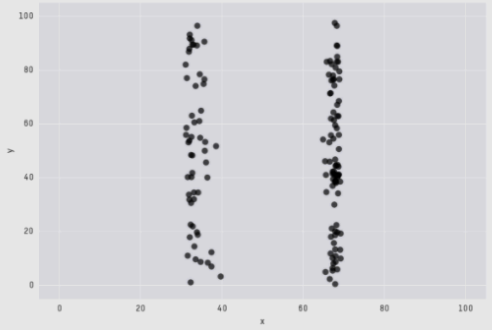
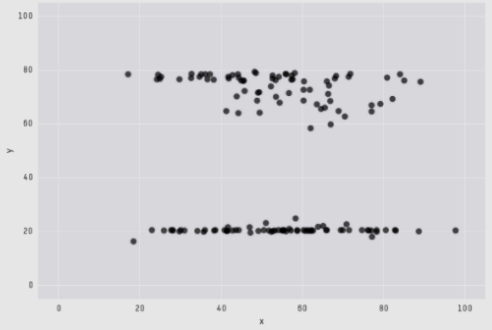
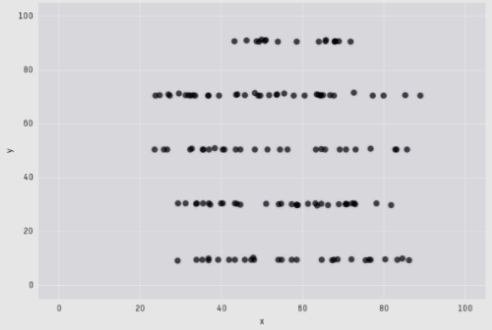
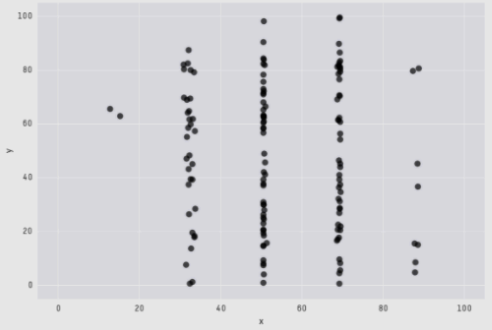
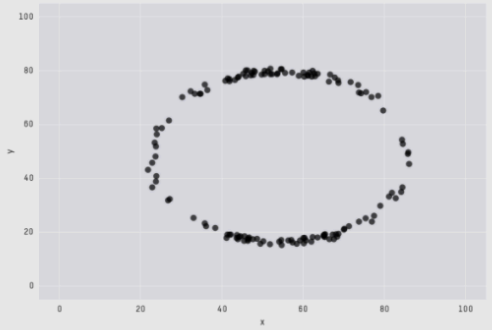
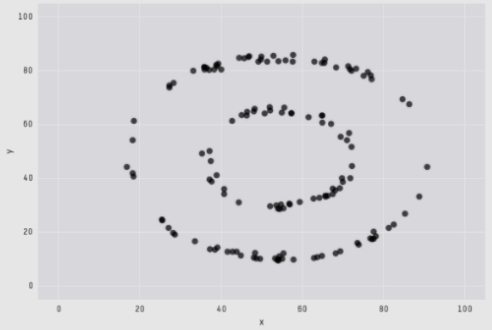
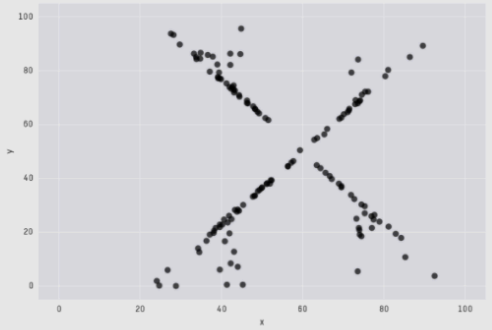
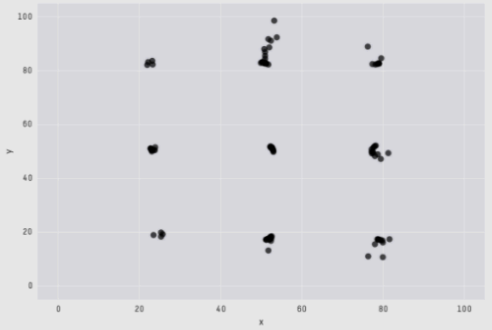
±0.1–0.3	Modest
±0.3–0.5	Moderate
±0.5–0.8	Strong
±0.8–0.9	Very strong



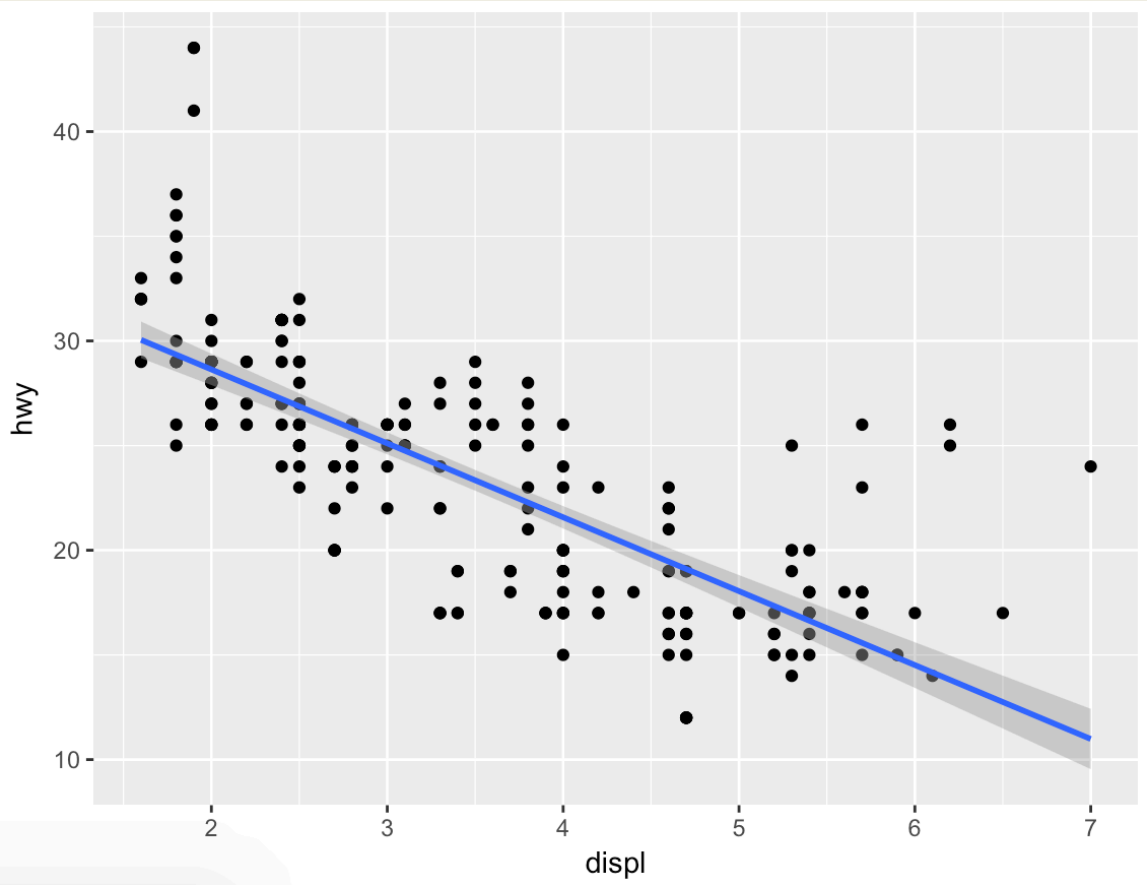
Visualizing correlation is, um, easy!

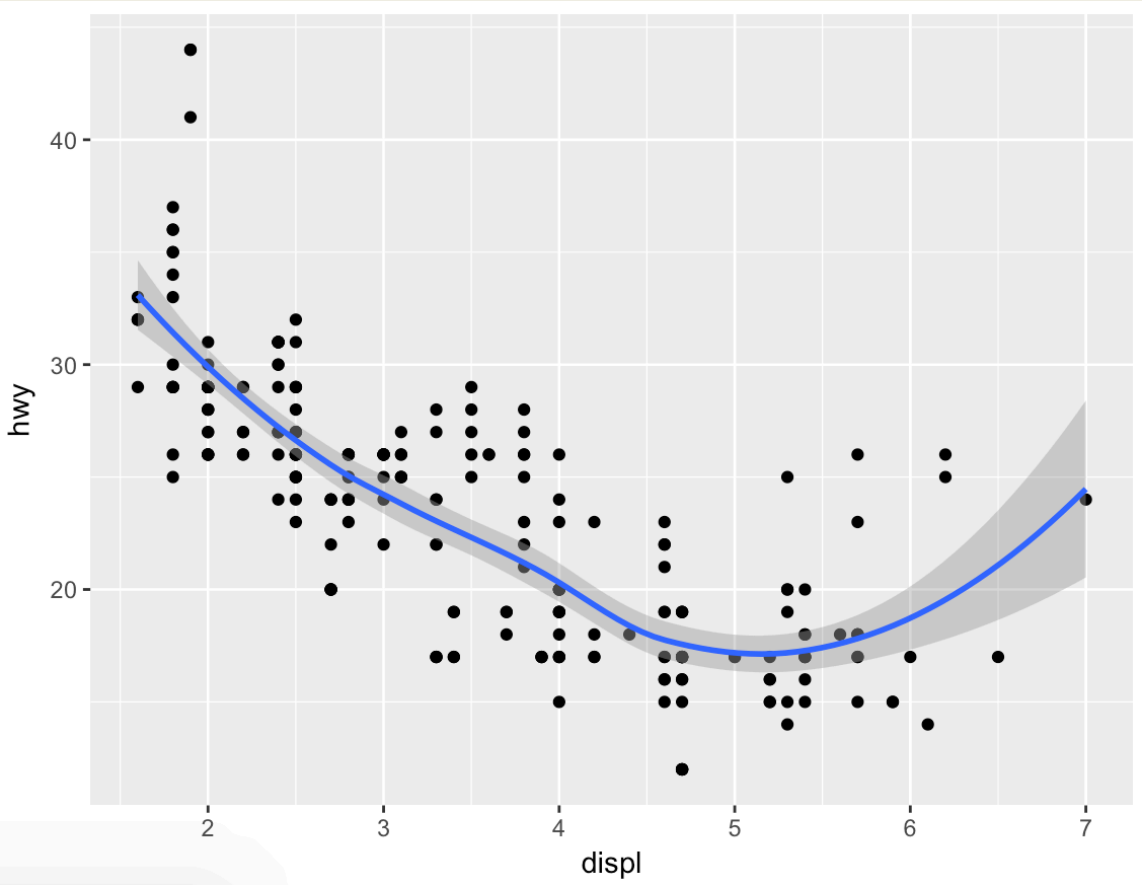
guessthecorrelation.com

X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

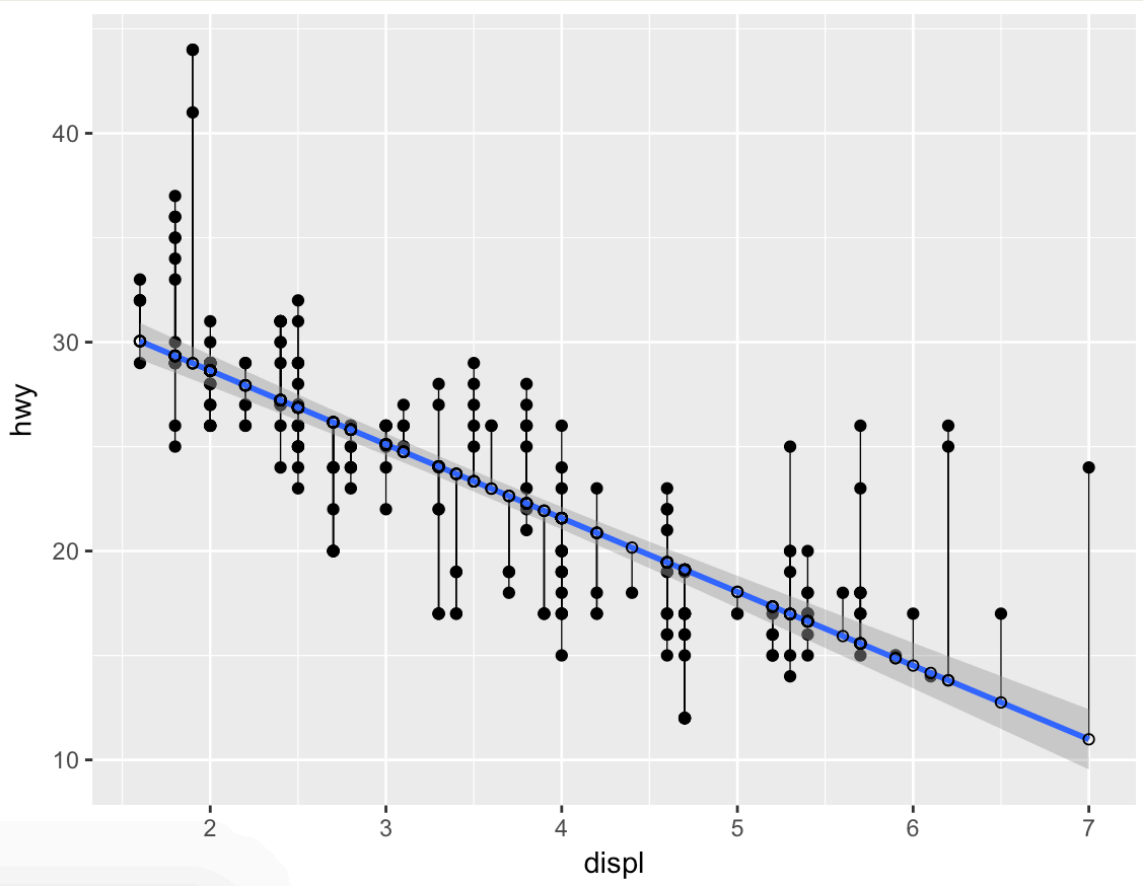













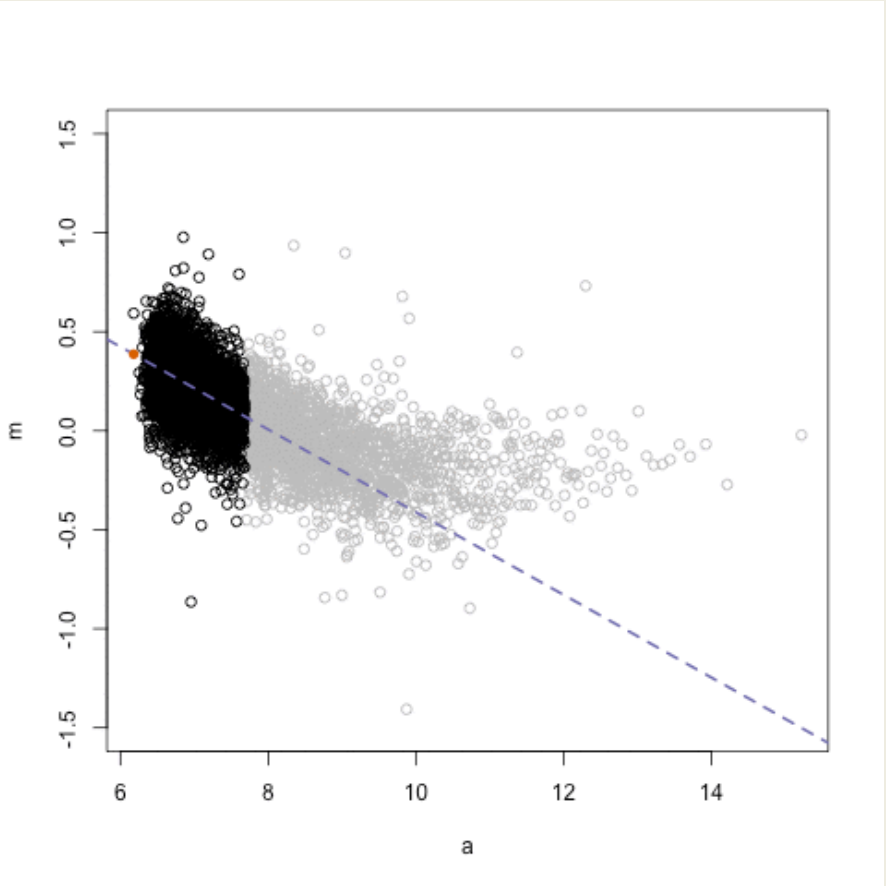
Drawing lines





Drawing lines



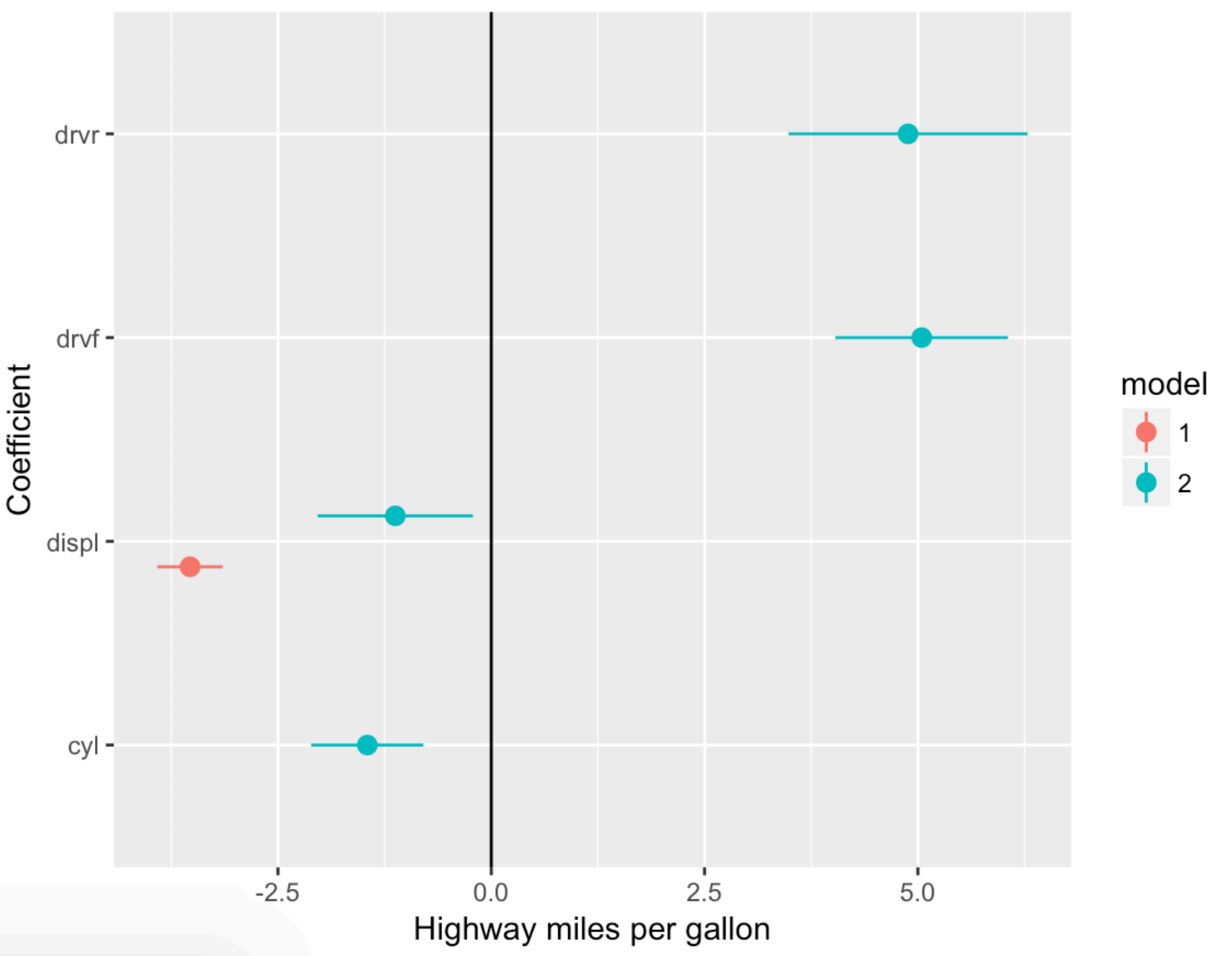


$$y = mx + b$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

```
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
```


drvr
drvf
displ
cyl
Coefficient
-2.5
0.0
2.5
5.0
Highway miles per gallon
model
1
2

🎉 R time again! 🎉