Data Analysis for Public Policy

(Tufte e-book + other topics)

The use of statistical methods to analyze data does not make a study any more “scientific,” “rigorous,” or “objective.”

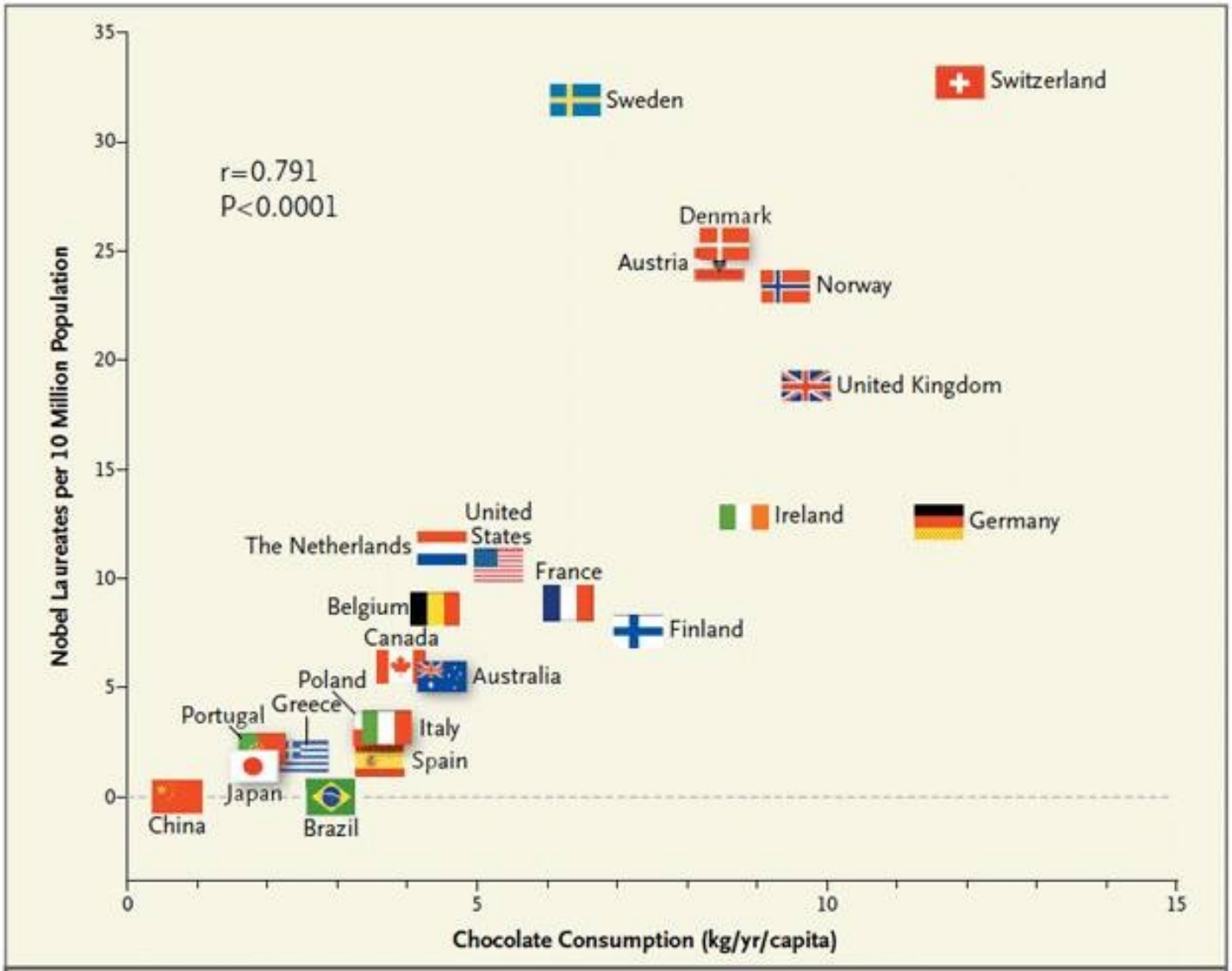
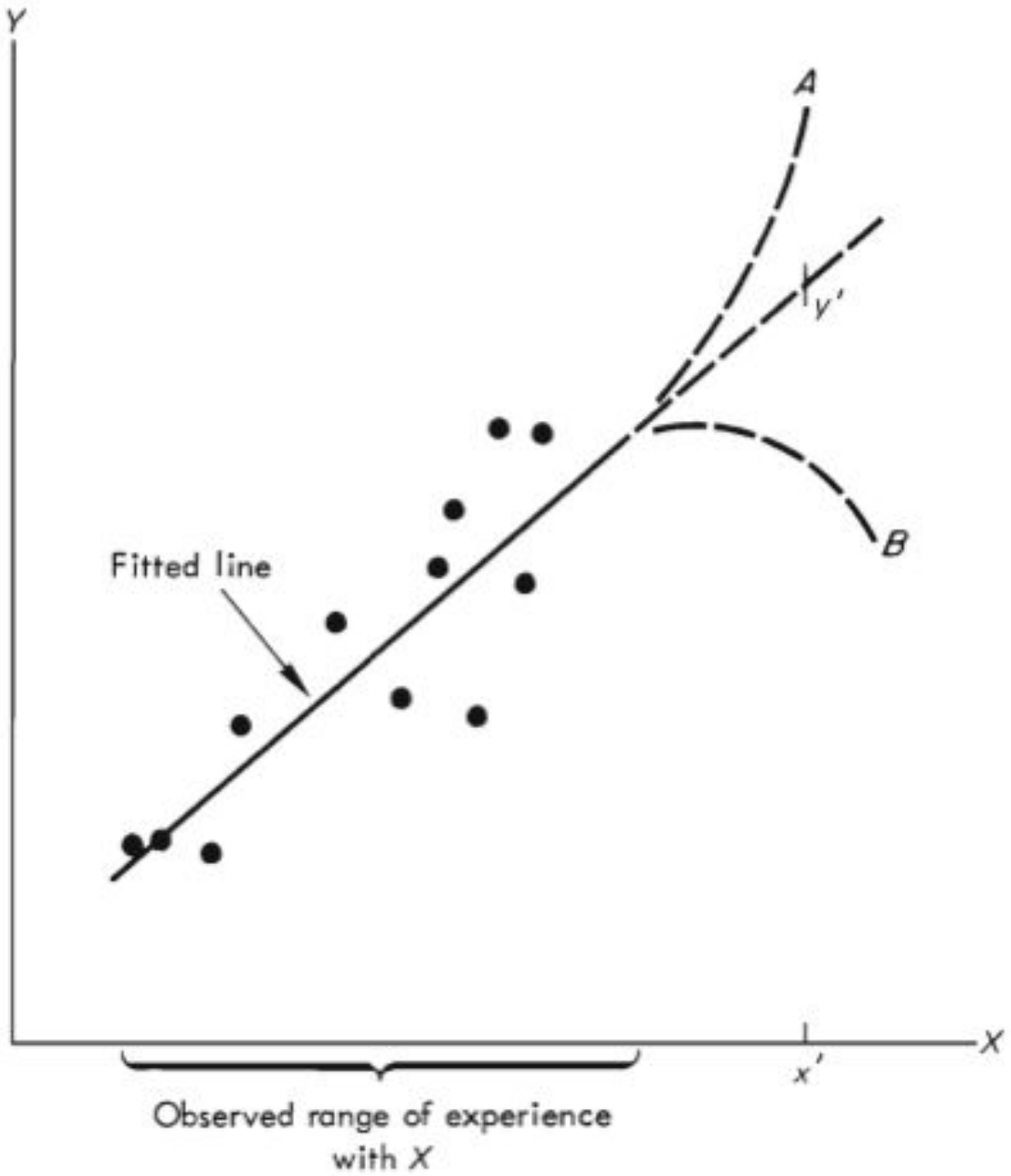


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

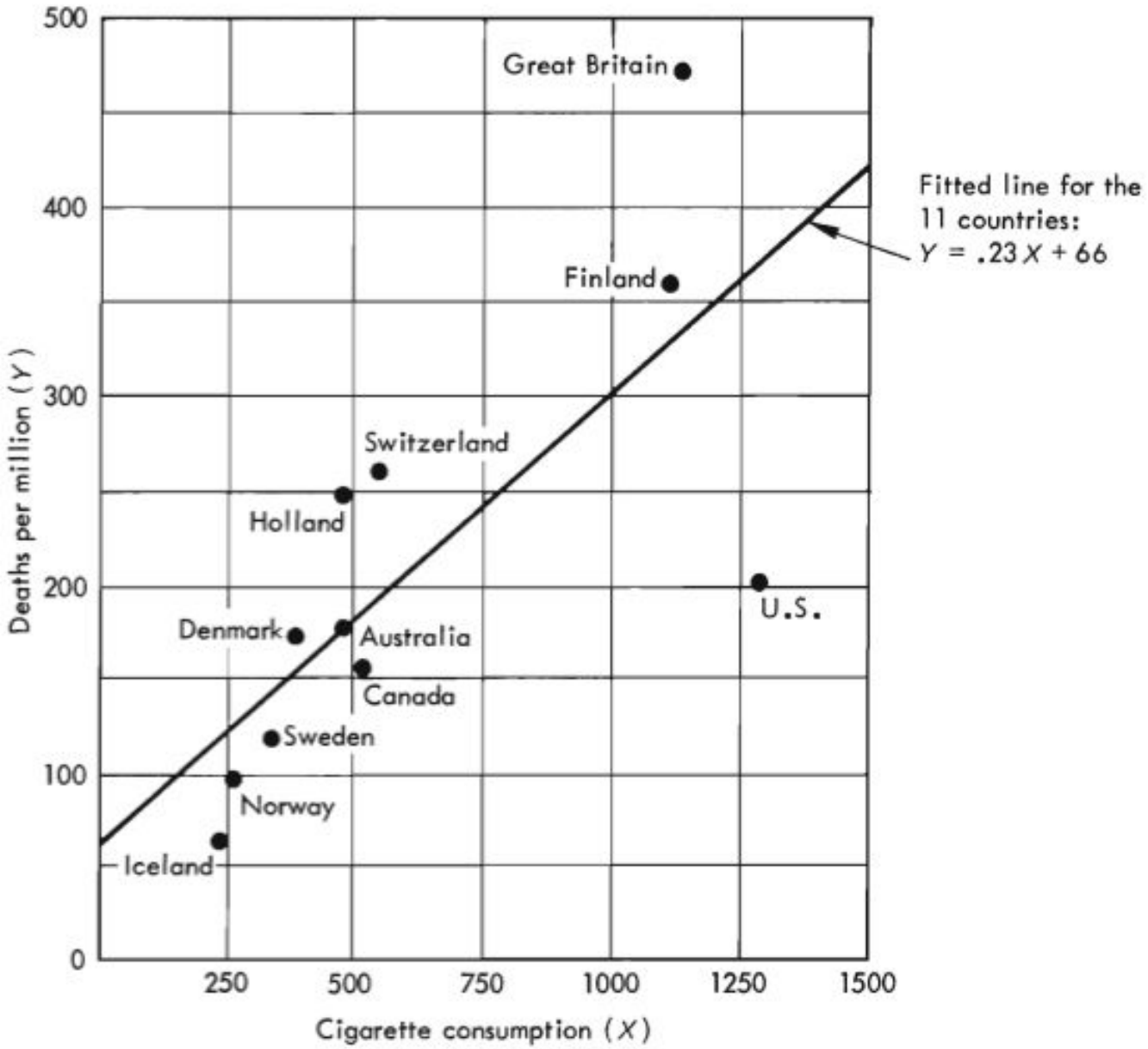
“It's tough to make predictions, especially about the future.”
Yogi Berra


Y
A
$Y_{x'}$
B
Fitted line
X
x'
Observed range of experience
with X

Research Design Defects

1. Correlation vs. causation

(or, failure to take into account the effects of other variables)


500
400
300
200
100
0
Deaths per million (Y)
250
500
750
1000
1250
1500
Cigarette consumption (X)
Great Britain
Finland
Switzerland
Holland
Denmark
Australia
Canada
Sweden
Norway
Iceland
U.S.
Fitted line for the
11 countries:
Y = .23X + 66

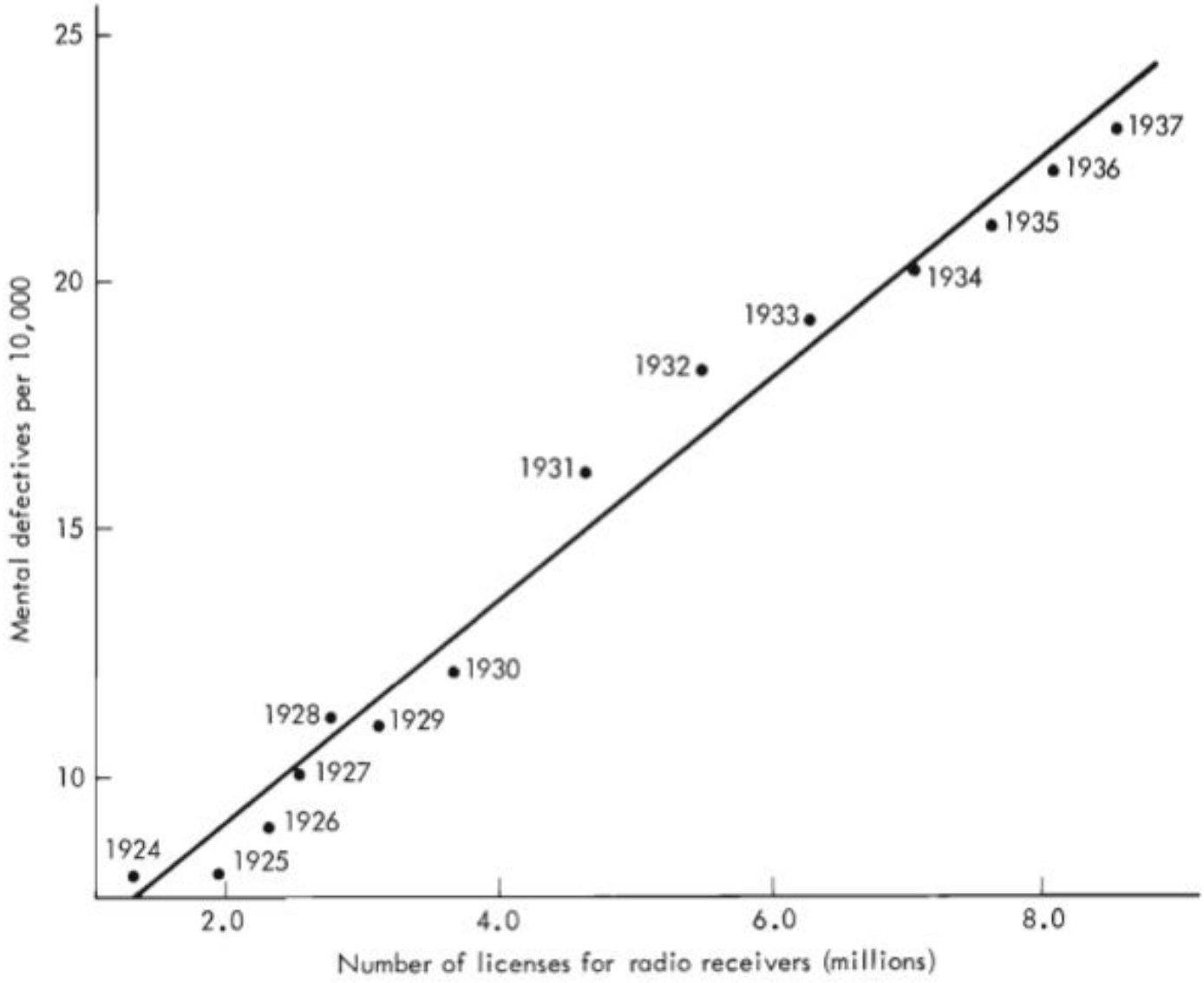


FIGURE 3-12 Radio receivers and mental defectives

$R^2 = 0.89$

Research Design Defects

1. Correlation vs. causation
(or, failure to take into account the effects of other variables)

2. Regression fallacy

Regression Toward the Mean: How Prior Selection Affects the Measurement of Future Performance

Consider the defects in research design in the following example:

> Students in a statistics course who needed remedial teaching (as indicated by their performance in the lower quartile of an achievement test in arithmetic) were assigned to a special class in sensitivity training. Soon the teacher of the special class was able to go into full-time educational consulting because of the success of his new book, *Ending Educational Hangups in Statistics: How Empathy Pays Off*. The book showed that the special class was strikingly effective because when the students in the special class took the tests again after only six months, their test scores had greatly increased—increased, in fact, almost all the way up to the average of the first test scores of all the students who initially took the arithmetic test.

1. Average "gain" for special class equals

$$\left(\begin{array}{l}\text{average of scores on}\\ \text{second test for special}\\ \text{class}\end{array}\right) \text{ minus } \left(\begin{array}{l}\text{average of scores on}\\ \text{first test for special}\\ \text{class}\end{array}\right)$$

2. "Improvement" relative to rest of class equals

$$\left(\begin{array}{l}\text{average of scores on}\\ \text{second test for special}\\ \text{group}\end{array}\right) \text{ minus } \left(\begin{array}{l}\text{average of scores for}\\ \text{whole class on the first}\\ \text{test}\end{array}\right)$$

What is $\mathbf{R^2}$?

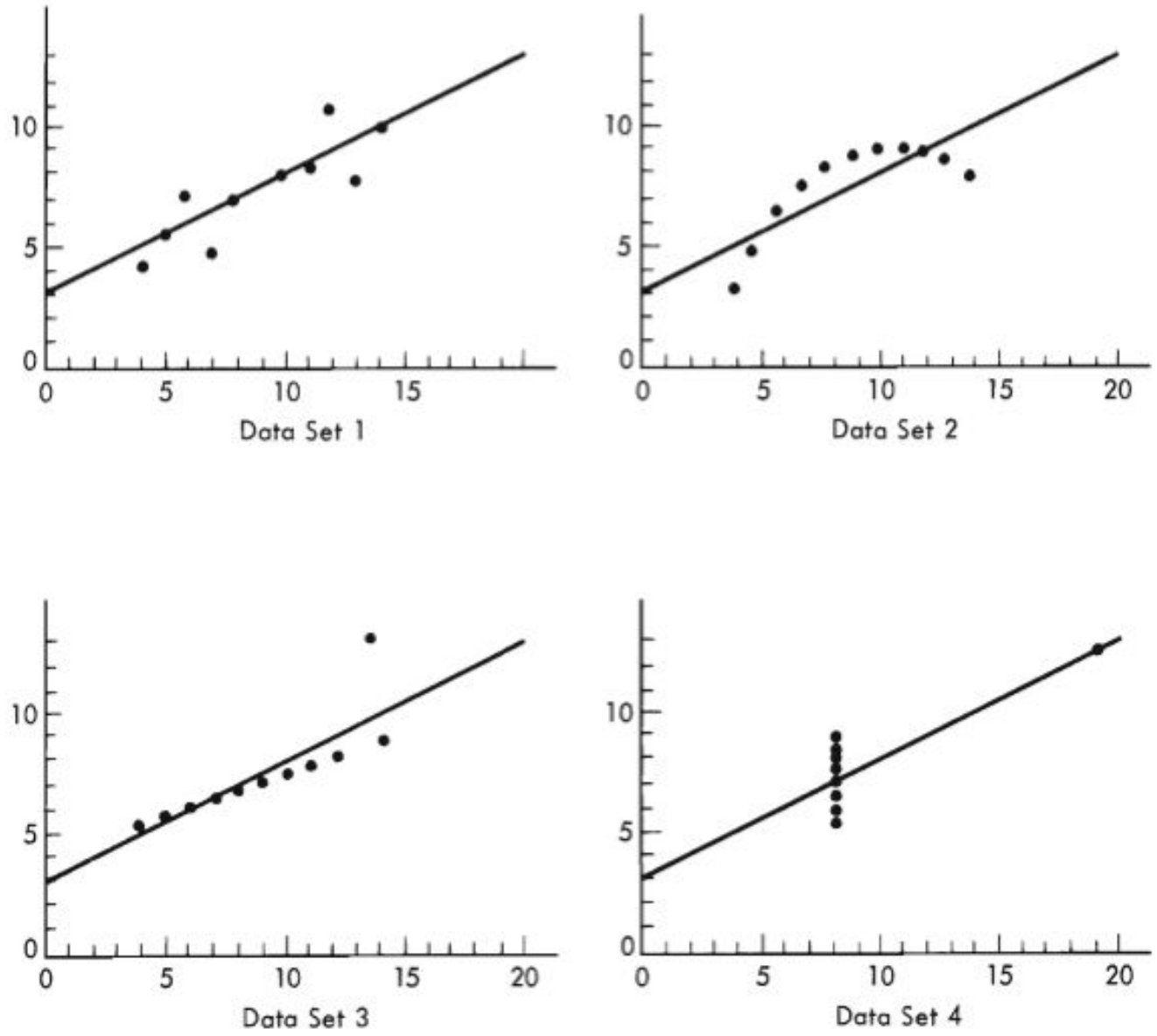


FIGURE 3-29 Scatterplots for the four data sets of Table 3-10
SOURCE: F. J. Anscombe, *op cit.*

What is **multicollinearity**?

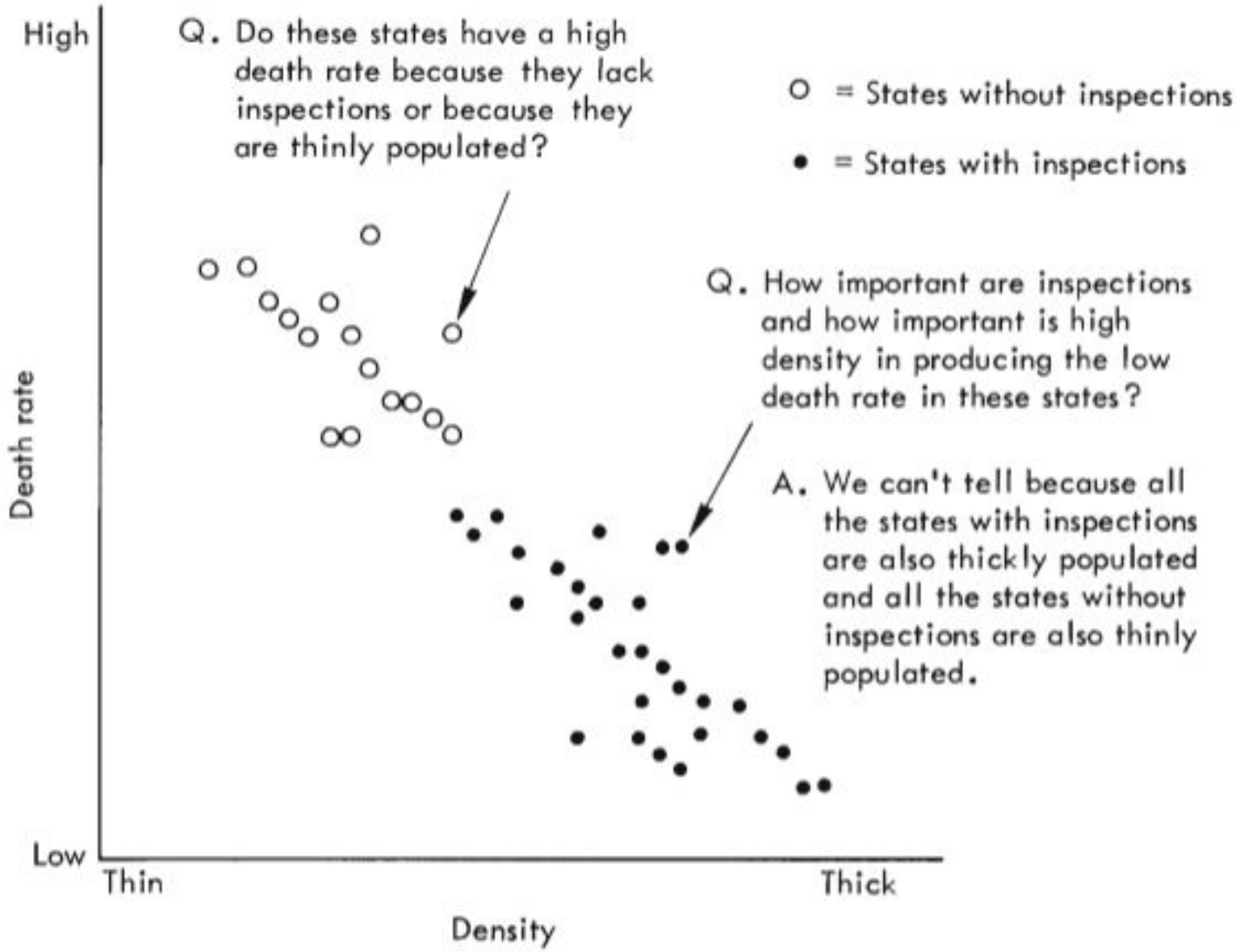


FIGURE 4-1 Hypothetical data showing collinearity between density and inspections

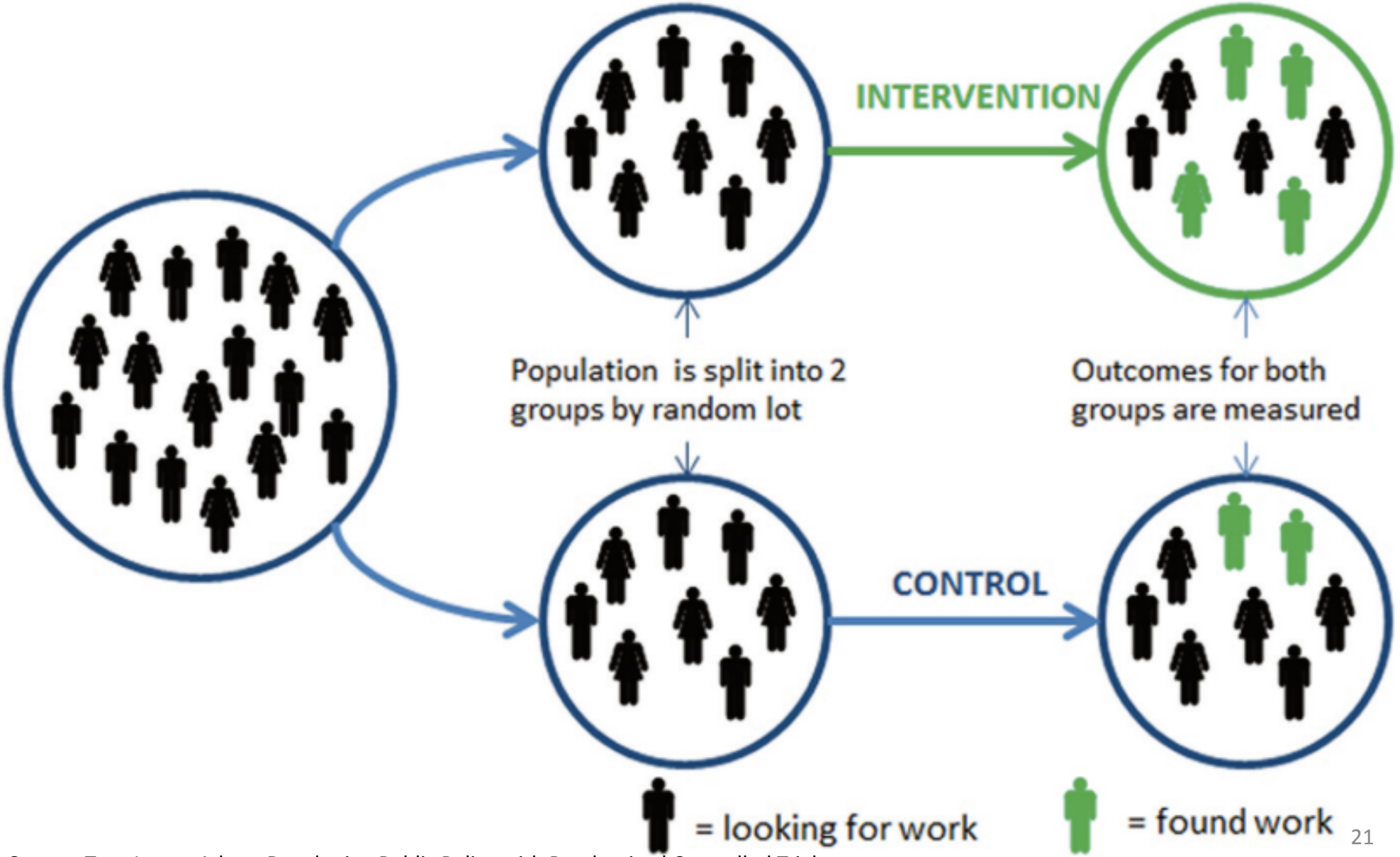
1. high intercorrelations between the describing variables,
2. large variances in the estimates of the regression coefficients,
3. large $R^2_{X_i}$,
4. large R^2_Y coupled with statistically nonsignificant regression coefficients,
5. large changes in the values of estimated regression coefficients when new variables are added to the regression, and
6. inability of computer program to compute regression coefficients (which occurs only in very severe cases of multicollinearity—in most cases the estimation procedures produce numbers as usual).

What is a **partial slope**?

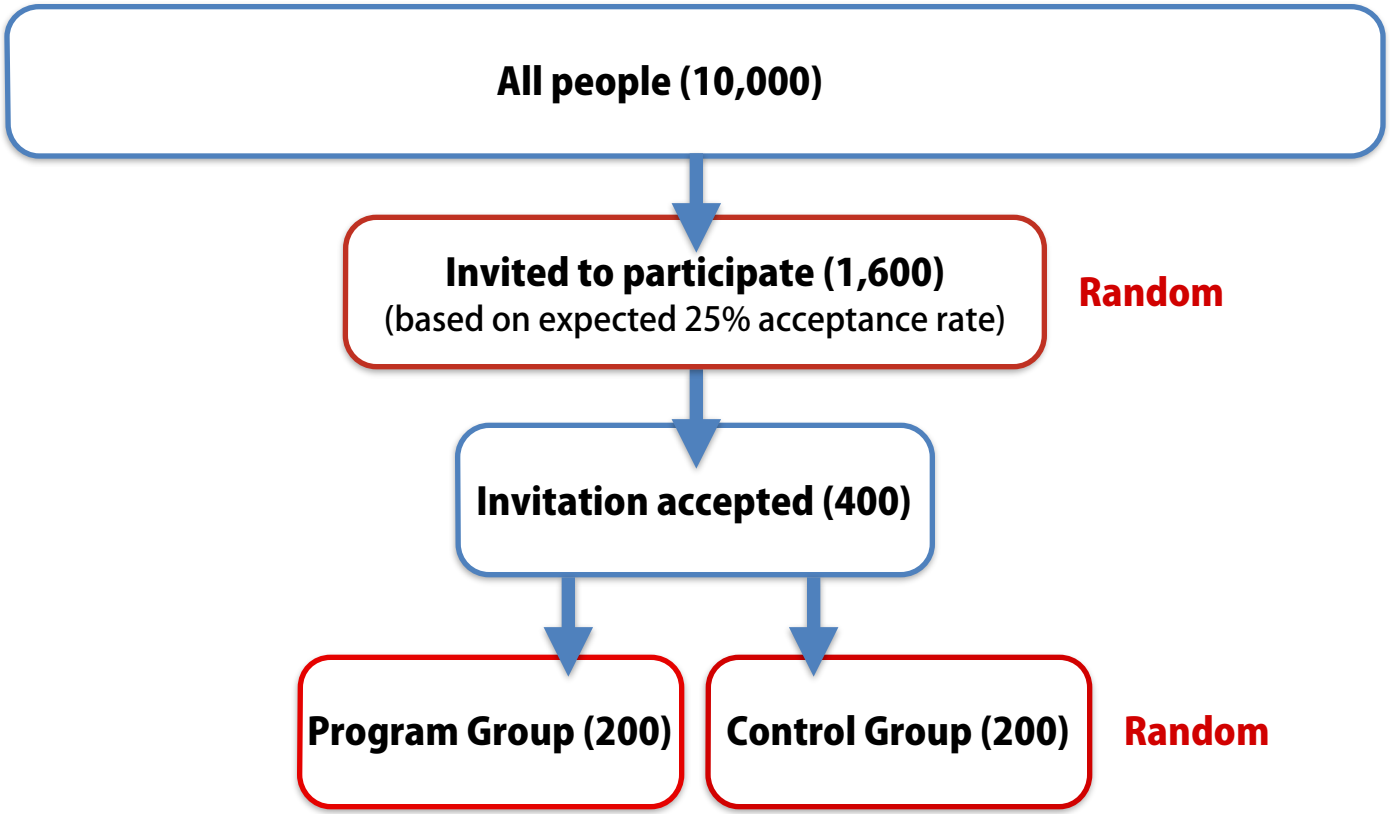
```
Call:
lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +
    HS.Grad + Frost + Area + Density, data = st)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47514 -0.45887 -0.06352  0.59362  1.21823

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.995e+01  1.843e+00  37.956  < 2e-16
Population   6.480e-05  3.001e-05   2.159   0.0367
Income       2.701e-04  3.087e-04   0.875   0.3867
Illiteracy   3.029e-01  4.024e-01   0.753   0.4559
Murder      -3.286e-01  4.941e-02  -6.652 5.12e-08
HS.Grad      4.291e-02  2.332e-02   1.840   0.0730
Frost       -4.580e-03  3.189e-03  -1.436   0.1585
Area        -1.558e-06  1.914e-06  -0.814   0.4205
Density     -1.105e-03  7.312e-04  -1.511   0.1385

Residual standard error: 0.7337 on 41 degrees of freedom
Multiple R-squared: 0.7501,     Adjusted R-squared: 0.7013
F-statistic: 15.38 on 8 and 41 DF,  p-value: 3.787e-10
```

`Population`:
population estimate as of July 1, 1975
`Income`:
per capita income (1974)
`Illiteracy`:
illiteracy (1970, percent of population)
`Life Exp`:
life expectancy in years (1969--71)
`Murder`:
murder and non-negligent manslaughter rate per 100,000 population (1976)
`HS Grad`:
percent high-school graduates (1970)
`Frost`:
mean number of days with minimum temperature below freezing
(1931--1960) in capital or large city
`Area`:
land area in square miles

other ways to use **randomness** in statistics



Source: Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials


All people (10,000)
Invited to participate (1,600)
(based on expected 25% acceptance rate)
Random
Invitation accepted (400)
Program Group (200)
Control Group (200)
Random

Evaluation of Accelerated Study in Associate Programs (ASAP) for Developmental Education Students

Table ES.1

Early Impacts on Selected Academic Outcomes

Outcome	Program Group	Control Group	Difference (Impact)	
First semester				
Enrolled[a] (%)	96.4	94.0	2.5	*
Full time	95.8	85.2	10.6	***
Part time	0.6	8.8	-8.1	***
Total credits attempted	16.1	13.9	2.2	***
College-level credits	10.5	10.3	0.2	
Developmental credits	5.6	3.6	2.0	***
Total credits earned	11.4	9.3	2.1	***
College-level credits	8.5	7.6	0.9	***
Developmental credits	2.9	1.7	1.1	***
Completed developmental requirements[b] (%)	46.6	31.9	14.7	***