Math 1070-2: Spring 2008

Lecture 2

Davar Khoshnevisan

Department of Mathematics
University of Utah
http://www.math.utah.edu/~davar

January 16, 2009

Recap on Computational Matters

- Data: 1, 2, 3, 4, 5, 6, 7, 8, 9
- Average (or mean)

$$\bar{x} = \frac{1+2+3+4+5+6+7+8+9}{9} = 5$$

- SD (standard deviation)

x	1	2	3	4	5	6	7	8	9
$x - \bar{x}$	-4	-3	-2	-1	0	1	2	3	4
$(x - \bar{x})^2$	16	9	4	1	0	1	4	9	16

- sum = $16 + 9 + 4 + 1 + 0 + 1 + 4 + 9 + 16 = 60$
- SD = $\sqrt{\text{sum}/(n-1)} = \sqrt{60/8} = \sqrt{7.5} \approx 2.74$
- Typical number in data: 5 give or take 2.74

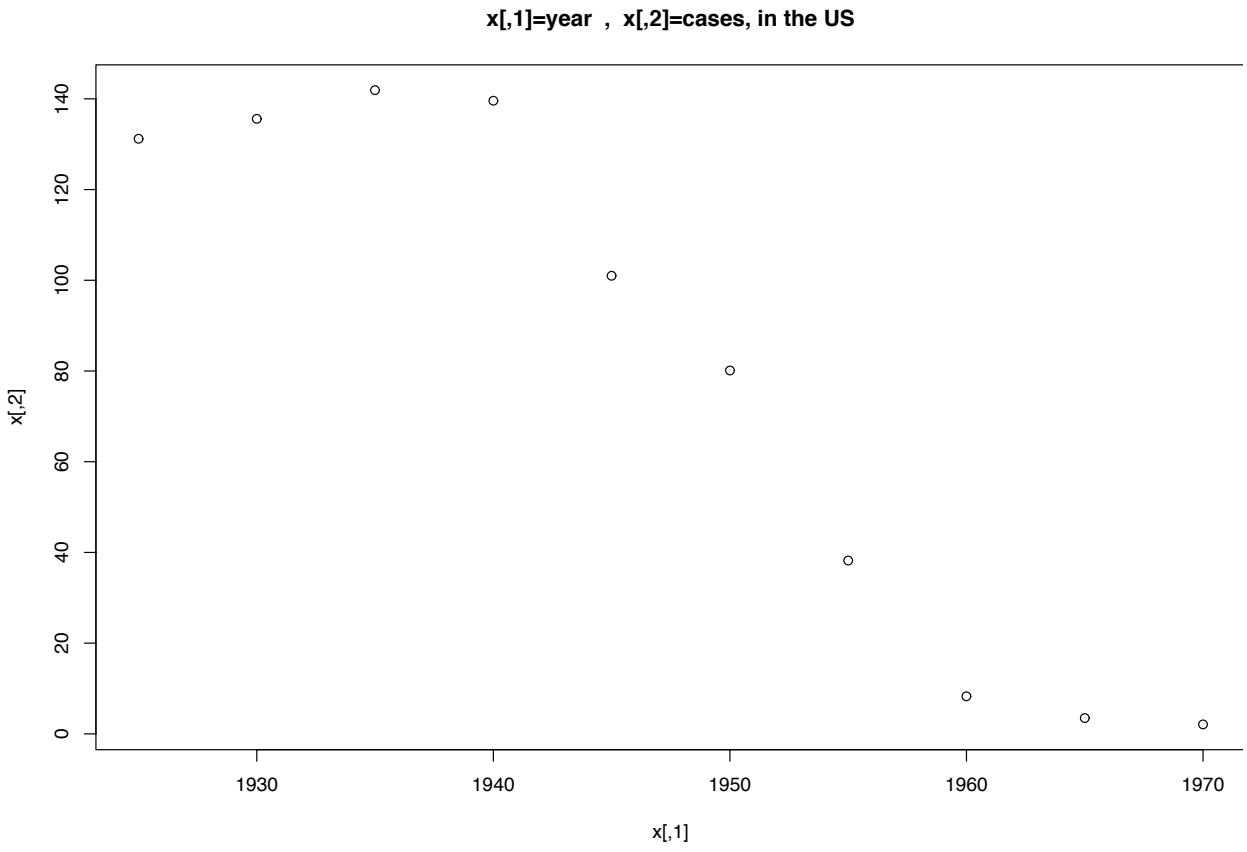
Association

Some Relevant Questions

- Does smoking cause cancer?
- Does more education mean less crime?
- Are newer drugs better than the older ones?
- ⋮
- Common theme: Look at “explanatory variables” to predict “response variables”
- Which is which in the previous three examples?

Scatterplots

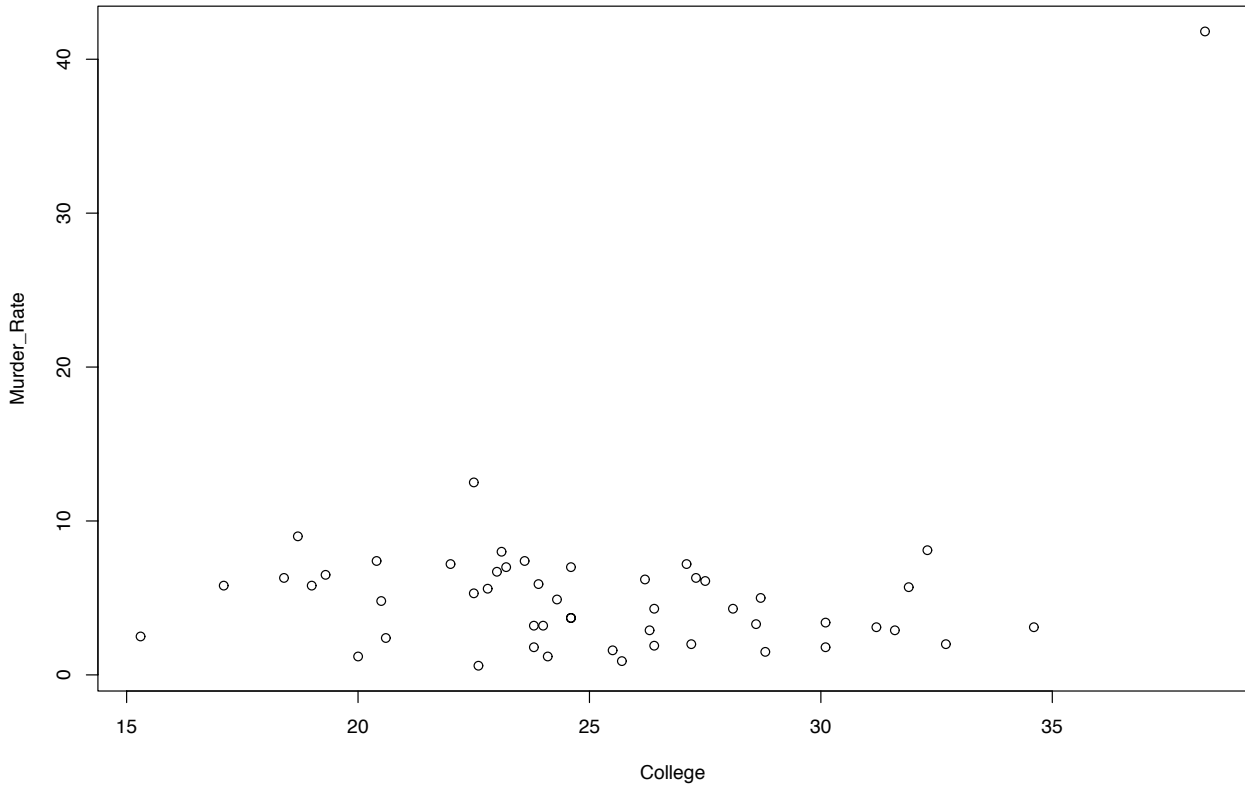
The Defeat of Whooping Cough in the U.S.



- Look for trends (1930-1940; 1940-1970)
- Negative association (strong? weak? linear relation?)

Murder Rates vs. College Education

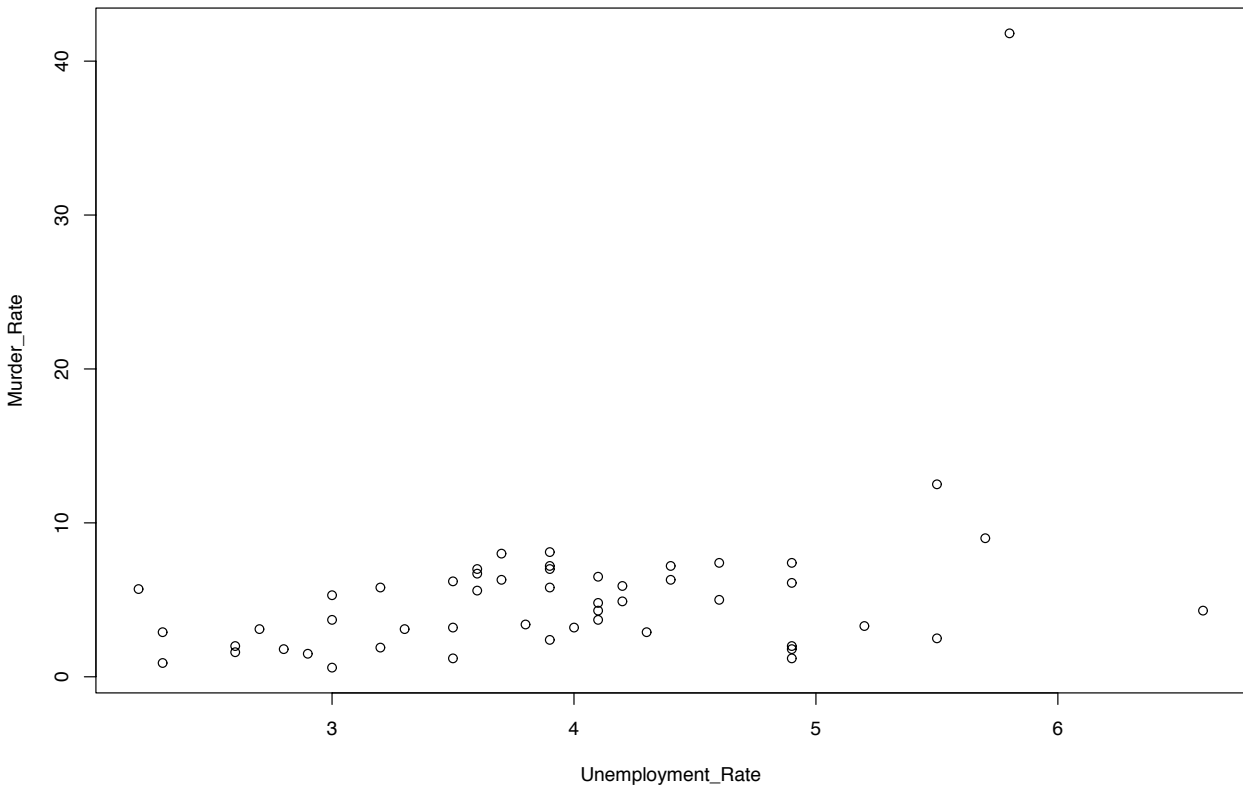
Life is not always so simple (Discussion)



- x-axis = percentage of college-educated people in that state
- y-axis = murder rate in that state

Murder Rates vs. Unemployment Rate

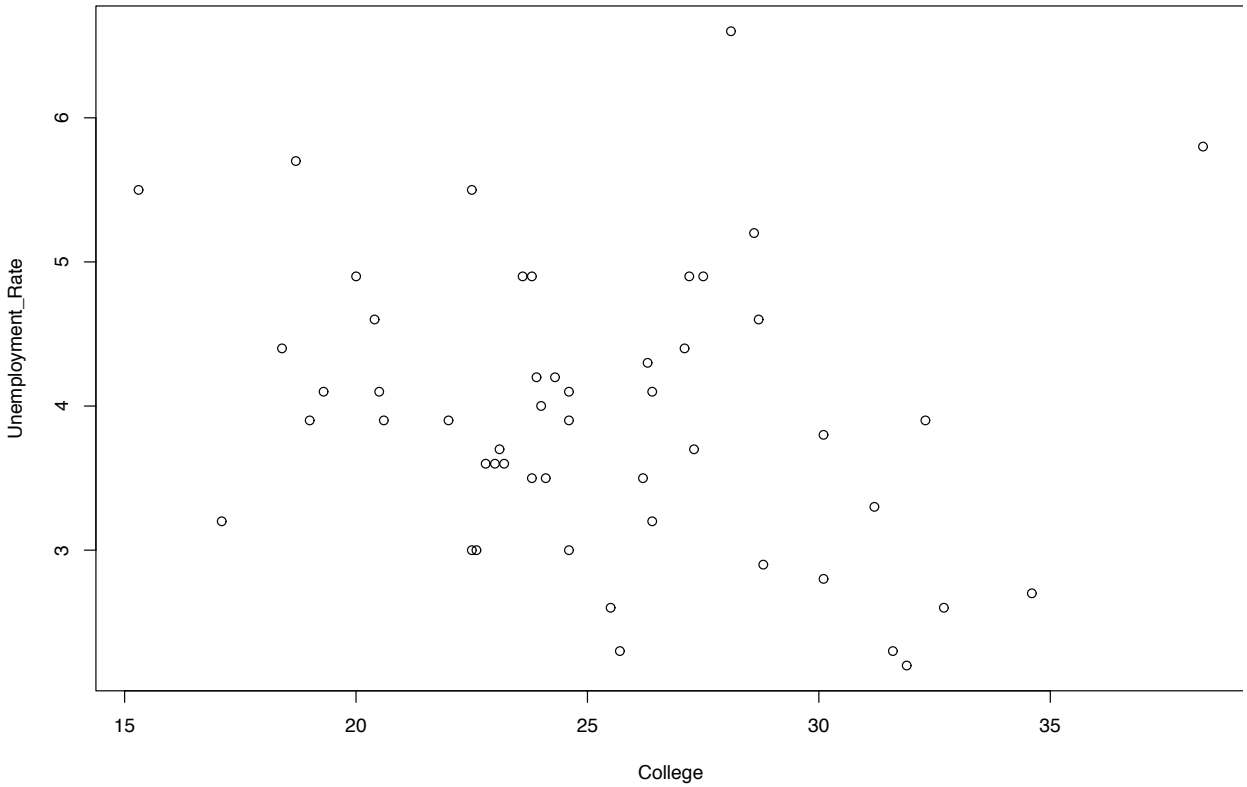
Discussion



- x-axis = percentage of unemployed people in that state
- y-axis = murder rate in that state

Higher Ed vs. Unemployment Rate

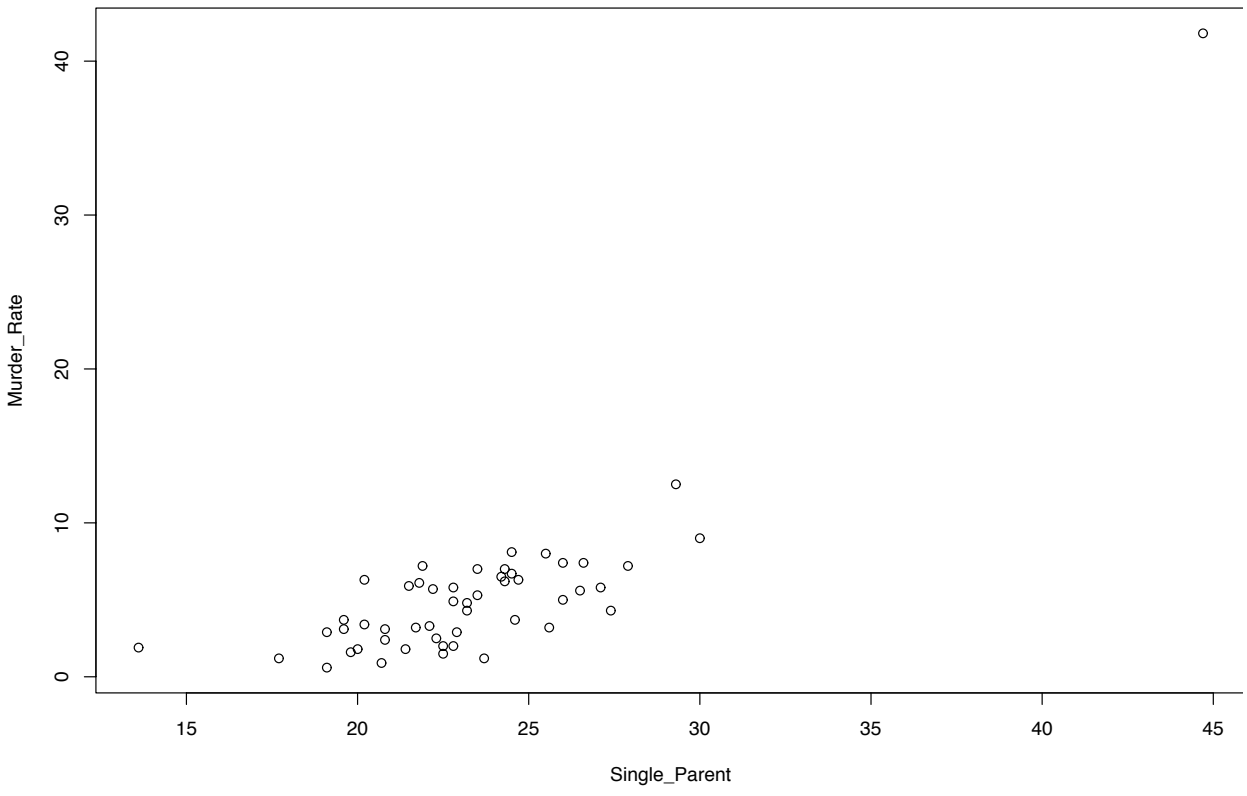
Discussion



- x-axis = percentage of college-educated people in that state
- y-axis = percentage of unemployed people in that state

Now Some Answers

Discussion



- x-axis = percentage of single parents in that state
- y-axis = murder rate in that state

Understanding Correlation (r)

- $-1 \leq r \leq 1$
- If $r \approx -1$ then strong negative association
- If $r \approx +1$ then strong positive association
- If $r \approx 0$ then no (or weak) linear association
- Example: (year vs. whooping cough) $r \approx -0.943$
- Example: (Single-parent-rate vs. murder rate) $r \approx 0.847$
- Example (College vs. unemployment rate) $r \approx -0.21$ ☺

How Do We Calculate r?

- Data type: $x_1, \ldots, x_n$ (e.g., year); $y_1, \ldots, y_n$ (e.g., no. of whooping-cough incidents)
- First Standardize you data:
  - $z_{x_i} = (x_i - \bar{x})/\mathrm{SD}_x$ (x_i in standard units)
  - $z_{y_i} = (y_i - \bar{y})/\mathrm{SD}_y$ (y_i in standard units)
- Then you compute:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}.$$

- Question: If $x_1 = y_1 \ldots x_n = y_n$, then what is r?

Standard Units

- Recall $z_{x_i} = (x_i - \bar{x})/\text{SD}_x$.
- Example: $\bar{x} = 5$, $\text{SD} = 2$
- $x = 4$ is $(4 - 5)/2 = -0.5$ standard units
- Interpretation: 0.5 SD's below $\bar{x}$ (verify)
- An advantage of thinking in standard units: They are absolute, unit-free numbers
- Not so helpful: I scored 10 points above average. (Out of how many points? How did others do? ☹)
- More helpful: I scored 2 standard deviations above average ☺

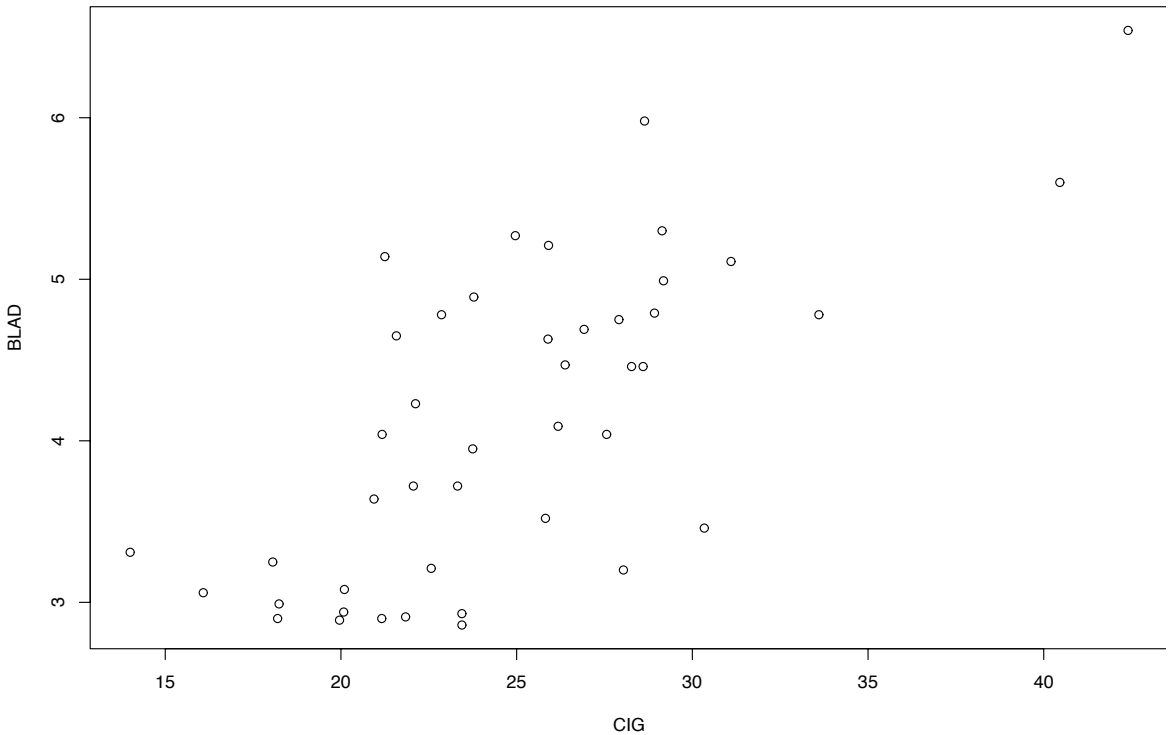
Cigarettes vs death by bladder cancer

Prediction, the Next Goal

- Basic problem: Have two quantitative variables (e.g., $x =$ no. of cigarettes smoked (heads/capita) versus $y =$ deaths per 100K population from bladder cancer) Does x affect y? How? Can we make predictions?
- Data from 1960 (by state)

Cigarettes vs death by bladder cancer

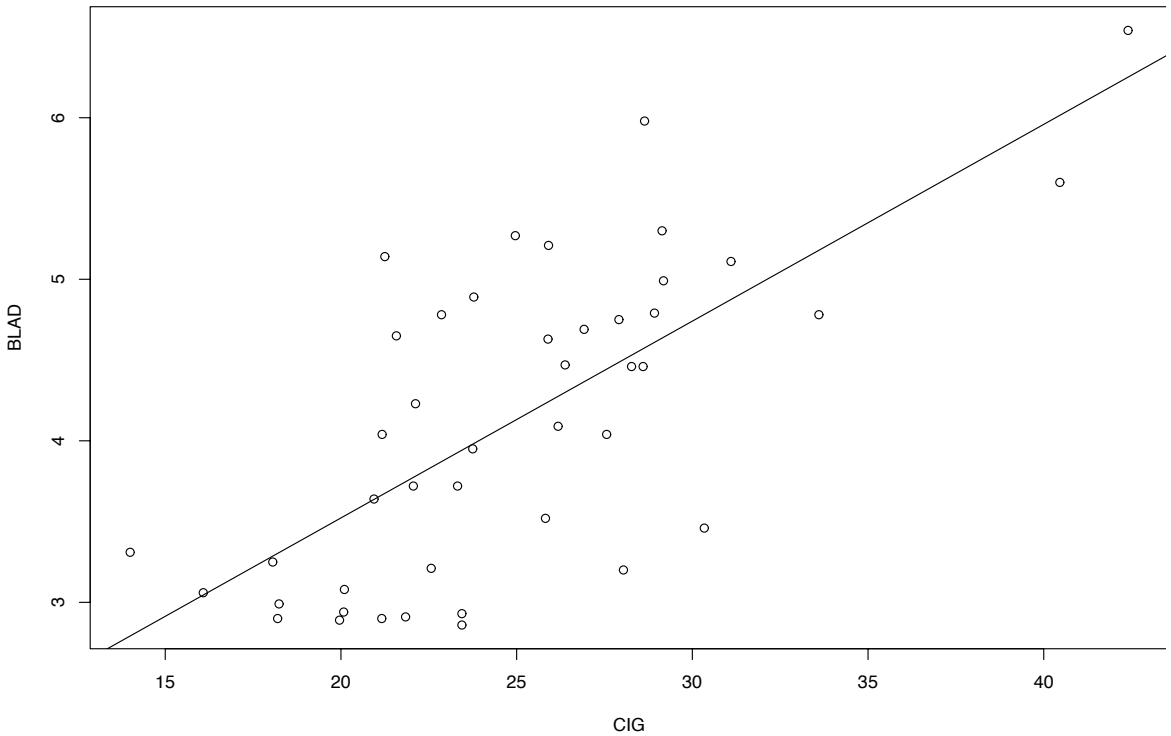
Prediction, the Next Goal



- $r \approx 0.7036219$
- Want a line of fit (simple-but-good description/prediction)

Cigarettes vs death by bladder cancer

Prediction, the Next Goal



- $r \approx 0.7036219$
- Want a line of fit (simple-but-good description/prediction)

Lines and plots

Blackboard lecture