Math 1070-2: Spring 2008

Lecture 3

Davar Khoshnevisan

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

January 23, 2009

Lines and plots

- Two numerical quantities: x [e.g., year] and y [e.g., income]
- Linear relationship:
$$y = a + bx.$$
- Set $x = 0 \;\rightarrow\; y = a$ [intercept]
- Set $y = 0 \;\rightarrow\; x = -a/b$ [abscissa]
- Move from x to $x' = x + \Delta \rightarrow$ move y to $y' = a + b(x + \Delta)$
- $\therefore$ a Δ-change $[x' - x]$ in the x value yields a $y' - y = b\Delta$-change in the y value
- $x' - x =$ run; $y' - y =$ rise
- $\therefore b =$ rise$/$run

Recall correlation (r)

- $-1 \leq r \leq 1$
- $r_{x,y} = r_{y,x}$
- 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 did 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 your 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:

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

- $\therefore r_{x,y} = r_{y,x}$
- Is it clear that $-1 \leq r \leq 1$? [Cauchy–Schwarz inequality]

[Simple] linear regression

- Two quantitative variables x [explanatory] and y [response]
- Sample: $(x_1, y_1), \ldots, (x_n, y_n)$
- Goal: Use the sample to find a "linear explanation" for y using x. That is, predict y with $\hat{y}$, where

$$\hat{y} = a + bx,$$

is the line that best fits the sample.

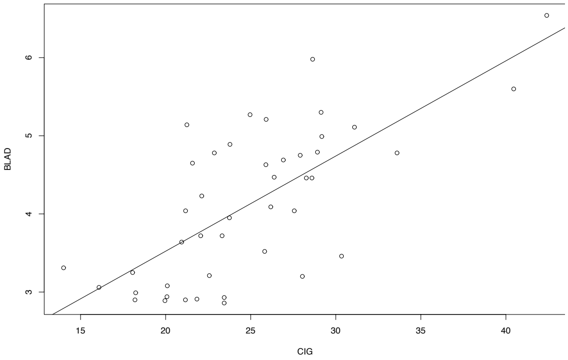
Cigarettes vs death by bladder cancer

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)

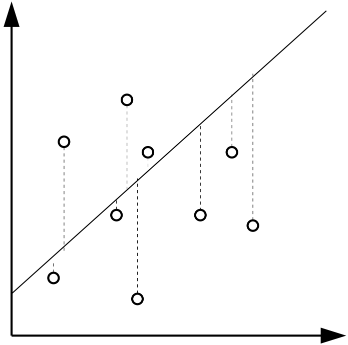
Cigarettes vs death by bladder cancer

Prediction, the Next Goal



- $r \approx 0.7036219$

The principle of least squares (LS)



- Slanted line = proposed line of fit
- circles = data
- length of dotted lines = residuals
- Goal: choose the line that minimizes $\sum(\text{residuals})^2$

The regression formula

- LS Fact: One SU-change in x is predicted by r SU-changes in $\hat{y}$
- $$\frac{\hat{y}-\bar{y}}{\mathrm{SD}_y} = r\left(\frac{x-\bar{x}}{\mathrm{SD}_x}\right)$$
- If $x=\bar{x}$ then $\hat{y}=\bar{y}$ (point of averages)
- We are not saying that
$$\frac{y-\bar{y}}{\mathrm{SD}_y} = r\left(\frac{x-\bar{x}}{\mathrm{SD}_x}\right)$$
- Can solve to obtain a formula:
$$\hat{y} = \overbrace{\bar{y} - \frac{r\,\mathrm{SD}_y}{\mathrm{SD}_x}}^{a} + \overbrace{\frac{r\,\mathrm{SD}_y}{\mathrm{SD}_x}}^{b} x$$

To remember the regression formula

- Either: $\text{SU}_{\hat{y}} = r\,\text{SU}_x$; or
- The regression line:
  - goes through $(\bar{x}\,,\bar{y})$—point of averages; and
  - has slope $r(\text{SD}_y/\text{SD}_x)$
- I will use the first memorization method
- You may use either, as long as you do it correctly

Education vs income

An example

- 1993 education/income data for 426 CA women (25–29 y.o.):
  - avg education ≈ 13 years; SD ≈ 3.1 years
  - avg income $\approx 17{,}500$ USD; SD $\approx 13{,}700$ USD
  - $r \approx 0.34$
- $x =$ education in years; $y =$ income in USD$\times 1{,}000$ USD
- Regression equation:

$$\frac{\hat{y} - 17.5}{13.7} = 0.34 \left(\frac{x - 13}{3.1} \right)$$

- If $x = 12$ years, then we predict income to be

$$\frac{\hat{y} - 17.5}{13.7} = 0.34 \left(\frac{12 - 13}{3.1} \right) \approx -0.1096774 \cdots \approx -0.1097$$

- Solve: $\hat{y} \approx (-0.1097 \times 13.7) + 17.5 \approx 15.997$ thousand USD

Femur length vs height

An example

- Goal: Given a found human femur (thighbone), predict the height
- Regression line: $\hat{y} = 61.4 + 2.4x$ [cm]
- Do the units count?
- If we find a femur that is 50 cm, then we predict the height to be $\hat{y} = 61.4 + 2.4(50) = 181.4$ cm
- Is every person with a fifty-cm femur 181.4 cm tall?
- What does this prediction mean?
- Interpolation vs. extrapolation

Batting averages

An example

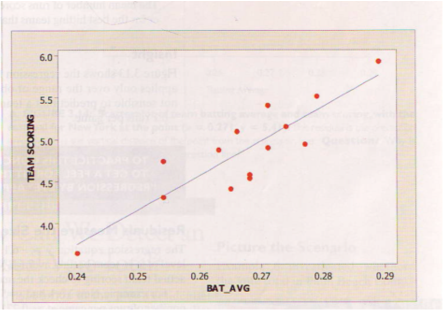
TABLE 3.5: Team Batting Average and Team Scoring (Mean Number of Runs Per Game) for American League Teams in 2003[a]

Team	Batting Average	Team Scoring
Boston	.289	5.9
Toronto	.279	5.5
Minnesota	.277	4.9
Kansas City	.274	5.2
Seattle	.271	4.9
New York	.271	5.4
Anaheim	.268	4.5
Baltimore	.268	4.6
Texas	.266	5.1
Tampa Bay	.265	4.4
Chicago	.263	4.9
Oakland	.254	4.7
Cleveland	.254	4.3
Detroit	.240	3.6

- x = batting avg; y = team score
- $\bar{x} \approx 0.267$; $\bar{y} = 4.85$
- $SD_x \approx 0.012$; $SD_y \approx 0.575$
- $r \approx 0.874$

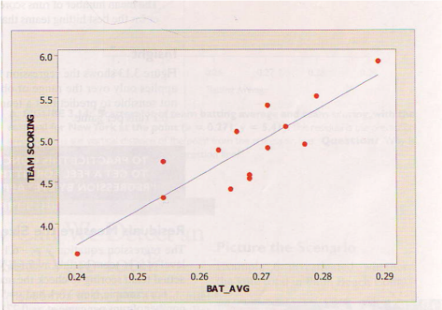
Batting averages

An example



- $\hat{y} = -6.1 + 41.2x$
- If $x = 0.278$ then $\hat{y} = -6.1 + 41.2(0.278) \approx 5.354$
- If $x = 0.268$ then $\hat{y} = -6.1 + 41.2(0.268) \approx 4.9416$
  [Baltimore? Anaheim?]

Extrapolation vs. interpolation



- (Interpolation) If $x = 0.278$ then $\hat{y} = -6.1 + 41.2(0.278) \approx 5.354$
- (Extrapolation) If $x = 0.1$ then $\hat{y} = -6.1 + 41.2(0.1) \approx -1.979$ ☹
- Extrapolate with care!!

Correlation vs causation

- Common error (a lot of firepersons at the scene $\not\to$ bigger fire)
- Can only determine causation by other methods; statistics might be used to corroborate
- One [serious] problem: Confounding [other, more important *lurking* variables]

Confounding

An example

- Should women take hormones, such as estrogen, after menopause?
- 1992 conclusion: “yes,” because women who took hormones reduced the risk of heart attacks by 30% to 50% [$\gg$ risk of taking hormones]
- Women who chose to take hormones are different than those who didn’t
  [richer; more educated; more frequent MD visits]
- 2002 conclusion: Hormones do not lower the risk
- The effect of the hormone(s) is confounded with the features of those who took hormones