# Introduction to Informatics

## Lecture 18: Inductive Model Building

Regression

Two statisticians were flying from L.A. to New York. About an hour into the flight, the pilot announced, "Unfortunately, we have lost an engine, but don't worry: There are three engines left. However, instead of five hours, it will take seven hours to get to New York.“

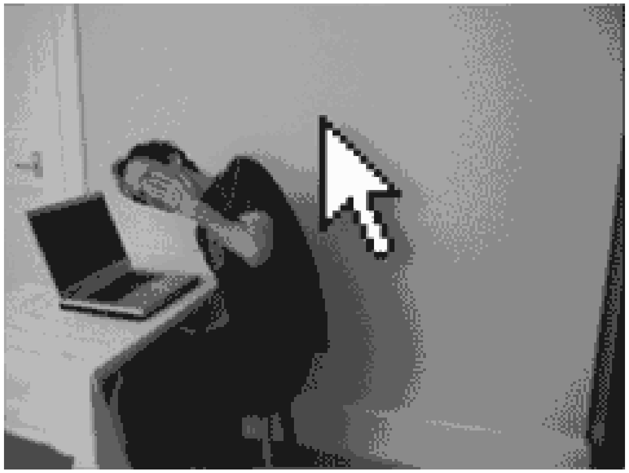
A little later, he told the passengers that a second engine had failed. "But we still have two engines left. We're still fine, except now it will take ten hours to get to New York."

Somewhat later, the pilot again came on the intercom and announced that a third engine had died. "But never fear, because this plane can fly on a single engine. Of course, it will now take 18 hours to get to New York."

At this point, one statistician turned to another and said, "Gee, I hope we don't lose that last engine, or we'll be up here forever!"

Luis M.Rocha and Santiago Schnell

# Readings until now



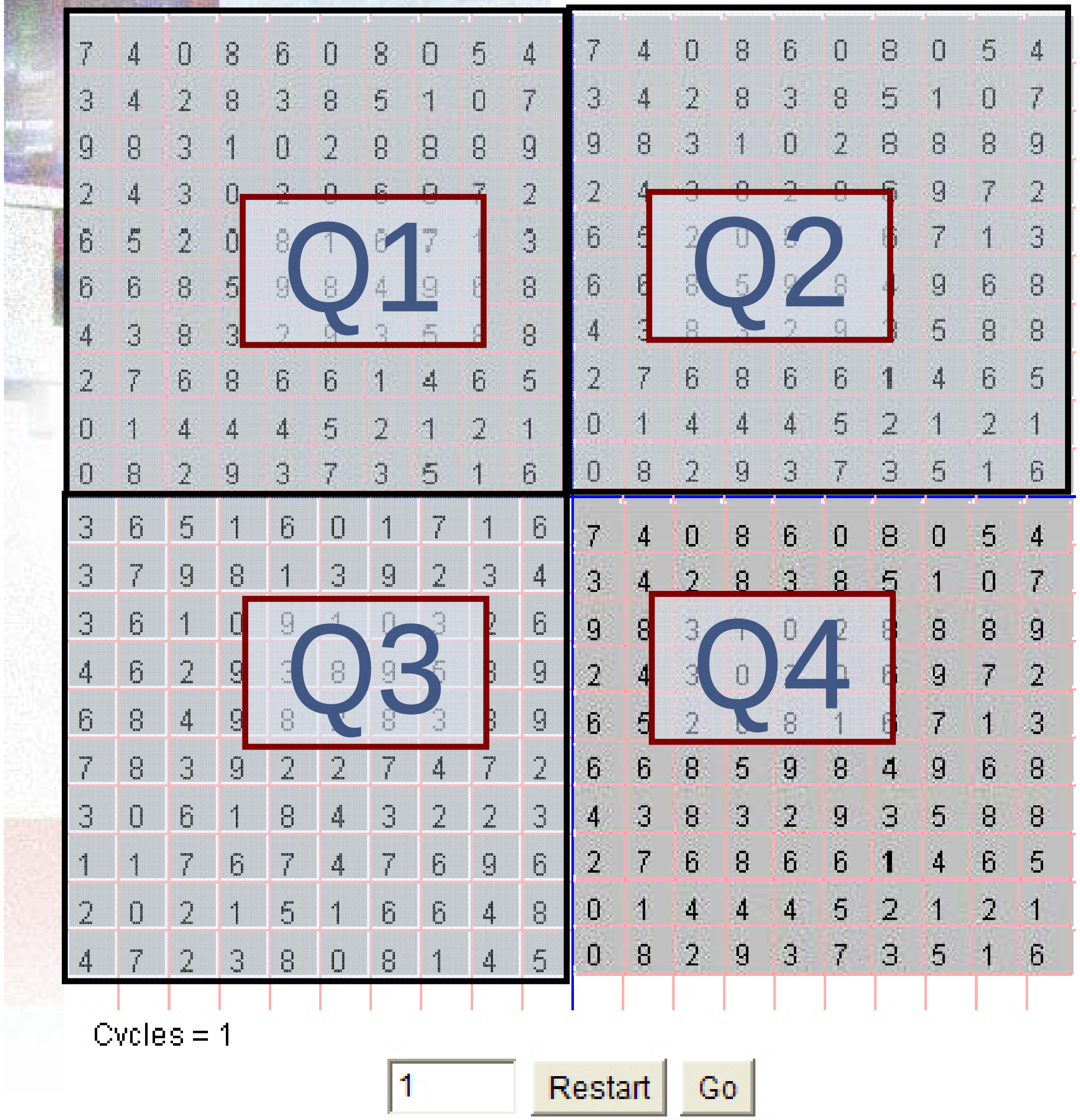
- Lecture notes
  - Posted online
    - http://informatics.indiana.edu/rocha/i101
      - *The Nature of Information*
      - *Technology*
      - *Modeling the World*
  - *@ infoport*
    - *http://infoport.blogspot.com*
  - From course package
    - Von Baeyer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.
      - Chapters 1, 4 (pages 1-12)
    - From  Andy Clark's book "*Natural-Born Cyborgs*"
      - Chapters 2 and 6 (pages 19 - 67)
    - From  Irv Englander's book "*The Architecture of Computer Hardware and Systems Software*"
      - Chapter 3: Data Formats (pp. 70-86)
    - Klir, J.G., U. St. Clair, and B.Yuan [1997]. Fuzzy Set Theory: foundations and Applications. Prentice Hall
      - Chapter 2: Classical Logic (pp. 87-97)
      - Chapter 3: Classical Set Theory (pp. 98-103)
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 105-129)
      - OPTIONAL: Chapter 4 (pages 131-136)
      - Chapter 13 (pages 147-155)

# Assignment Situation

- Labs
  - Past
    - Lab 1: Blogs
      - Closed (Friday, January 19): Grades Posted
    - Lab 2: Basic HTML
      - Closed (Wednesday, January 31): Grades Posted
    - Lab 3: Advanced HTML: Cascading Style Sheets
      - Closed (Friday, February 2): Grades Posted
    - Lab 4: More HTML and CSS
      - Closed (Friday, February 9): Grades Posted
    - Lab 5: Introduction to Operating Systems: Unix
      - Closed (Friday, February 16): Grades Posted
    - Lab 6: More Unix and FTP
      - Closed (Friday, February 23): Grades Posted
    - Lab 7: Logic Gates
      - Closed (Friday, March 9): Being Graded
  - Next: Lab 8
    - Intro to Statistical Analysis using Excel
      - March 22 & 23, Due Friday, March 30

- Assignments
  - Individual
    - First installment
      - Closed: February 9: Grades Posted
    - Second Installment
      - Past: March 2, Being Grades Posted
    - Third installment
      - Presented on March 8th, Due on March 30th
  - Group
    - First Installment
      - Past: March 9th, Being graded
    - Second Installment
      - March 29; Due Friday, April 6

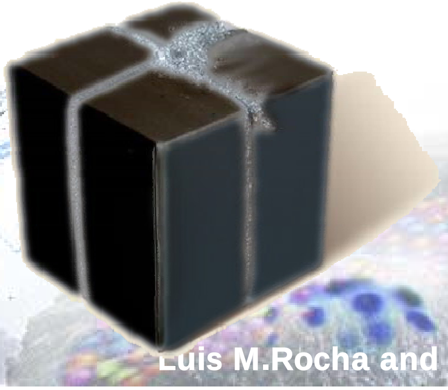
# Individual Assignment – Part III



| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 4 | 0 | 8 | 6 | 0 | 8 | 0 | 5 | 4 | 7 | 4 | 0 | 8 | 6 | 0 | 8 | 0 | 5 | 4 |
| 3 | 4 | 2 | 8 | 3 | 8 | 5 | 1 | 0 | 7 | 3 | 4 | 2 | 8 | 3 | 8 | 5 | 1 | 0 | 7 |
| 9 | 8 | 3 | 1 | 0 | 2 | 8 | 8 | 8 | 9 | 9 | 8 | 3 | 1 | 0 | 2 | 8 | 8 | 8 | 9 |
| 2 | 4 | 3 | 0 | 2 | 0 | 6 | 9 | 7 | 2 | 2 | 4 | 3 | 0 | 2 | 0 | 6 | 9 | 7 | 2 |
| 6 | 5 | 2 | 0 | 8 | 1 | 6 | 7 | 1 | 3 | 6 | 5 | 2 | 0 | 8 | 1 | 6 | 7 | 1 | 3 |
| 6 | 6 | 8 | 5 | 9 | 8 | 4 | 9 | 6 | 8 | 6 | 6 | 8 | 5 | 9 | 8 | 4 | 9 | 6 | 8 |
| 4 | 3 | 8 | 3 | 2 | 9 | 3 | 5 | 8 | 8 | 4 | 3 | 8 | 3 | 2 | 9 | 3 | 5 | 8 | 8 |
| 2 | 7 | 6 | 8 | 6 | 6 | 1 | 4 | 6 | 5 | 2 | 7 | 6 | 8 | 6 | 6 | 1 | 4 | 6 | 5 |
| 0 | 1 | 4 | 4 | 4 | 5 | 2 | 1 | 2 | 1 | 0 | 1 | 4 | 4 | 4 | 5 | 2 | 1 | 2 | 1 |
| 0 | 8 | 2 | 9 | 3 | 7 | 3 | 5 | 1 | 6 | 0 | 8 | 2 | 9 | 3 | 7 | 3 | 5 | 1 | 6 |
| 3 | 6 | 5 | 1 | 6 | 0 | 1 | 7 | 1 | 6 | 7 | 4 | 0 | 8 | 6 | 0 | 8 | 0 | 5 | 4 |
| 3 | 7 | 9 | 8 | 1 | 3 | 9 | 2 | 3 | 4 | 3 | 4 | 2 | 8 | 3 | 8 | 5 | 1 | 0 | 7 |
| 3 | 6 | 1 | 0 | 9 | 1 | 0 | 3 | 2 | 6 | 9 | 8 | 3 | 1 | 0 | 2 | 8 | 8 | 8 | 9 |
| 4 | 6 | 2 | 9 | 3 | 8 | 9 | 5 | 3 | 9 | 2 | 4 | 3 | 0 | 2 | 0 | 6 | 9 | 7 | 2 |
| 6 | 8 | 4 | 9 | 8 | [illegible] | 8 | 3 | 3 | 9 | 6 | 5 | 2 | 0 | 8 | 1 | 6 | 7 | 1 | 3 |
| 7 | 8 | 3 | 9 | 2 | 2 | 7 | 4 | 7 | 2 | 6 | 6 | 8 | 5 | 9 | 8 | 4 | 9 | 6 | 8 |
| 3 | 0 | 6 | 1 | 8 | 4 | 3 | 2 | 2 | 3 | 4 | 3 | 8 | 3 | 2 | 9 | 3 | 5 | 8 | 8 |
| 1 | 1 | 7 | 6 | 7 | 4 | 7 | 6 | 9 | 6 | 2 | 7 | 6 | 8 | 6 | 6 | 1 | 4 | 6 | 5 |
| 2 | 0 | 2 | 1 | 5 | 1 | 6 | 6 | 4 | 8 | 0 | 1 | 4 | 4 | 4 | 5 | 2 | 1 | 2 | 1 |
| 4 | 7 | 2 | 3 | 8 | 0 | 8 | 1 | 4 | 5 | 0 | 8 | 2 | 9 | 3 | 7 | 3 | 5 | 1 | 6 |

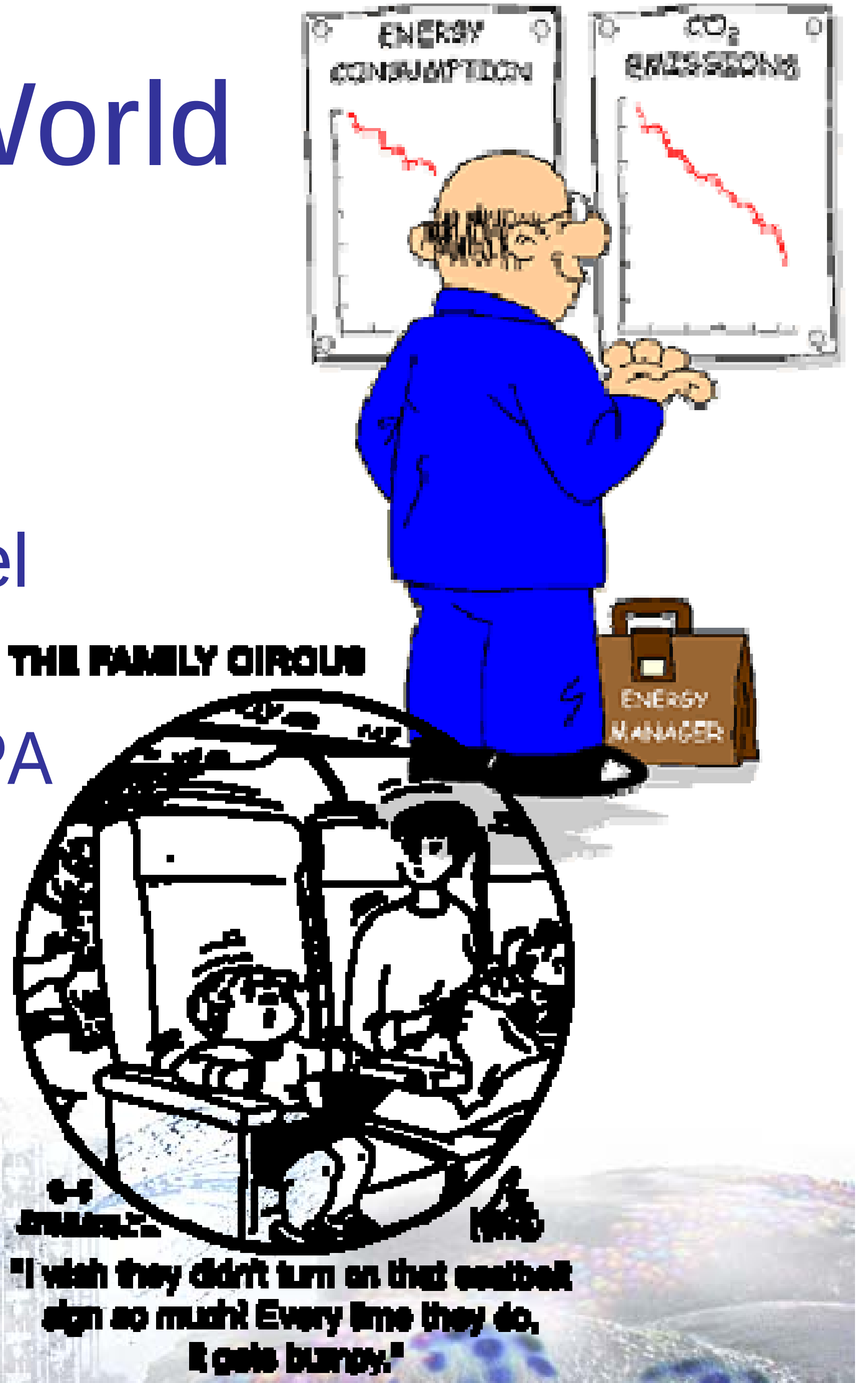
Cvcles = 1

- Step by step analysis of “dying” squares
  - 3rd Installment
    - Presented: March 8th
    - Due: March 30th
  - 4th Installment
    - Presented: April 5th
    - Due: April 20th
- Use descriptive statistics
  - To uncover rules inductively
    - E.g. the behavior of evens and odds, individual numbers, or ranges of cycles, etc.



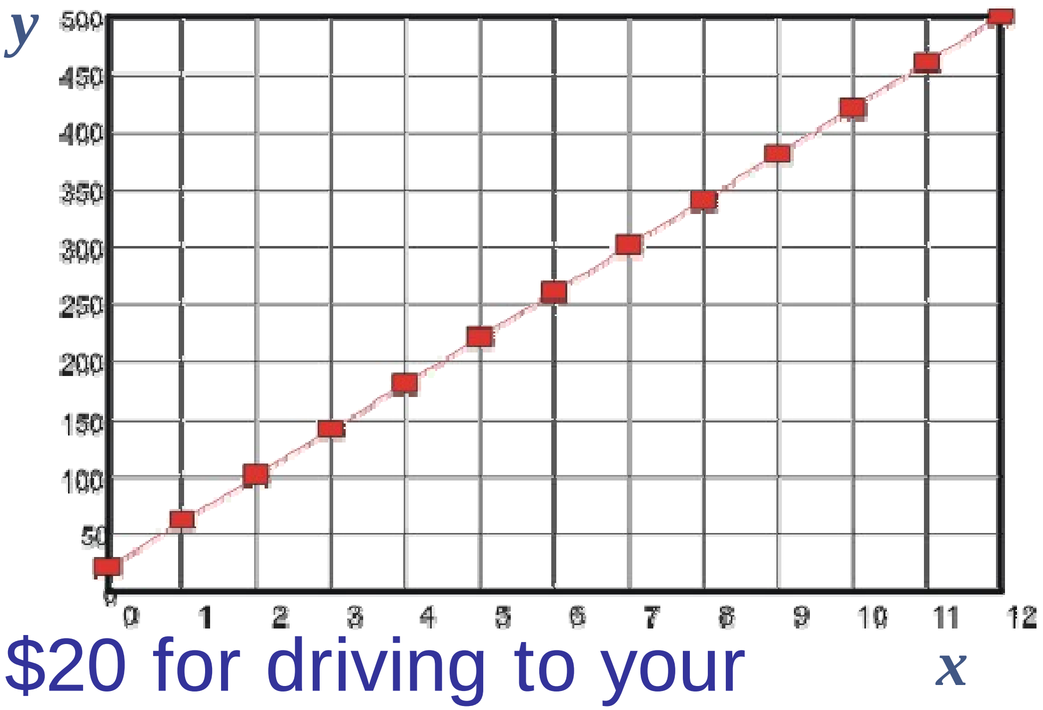
# Relations in the World

- Is there a relationship between two variables?
  - Years of schooling and level of income
  - High-school and college GPA
  - Inflation rate and prime lending rate
- What is the relationship?
  - Regression analysis



"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

# Linear Relationship



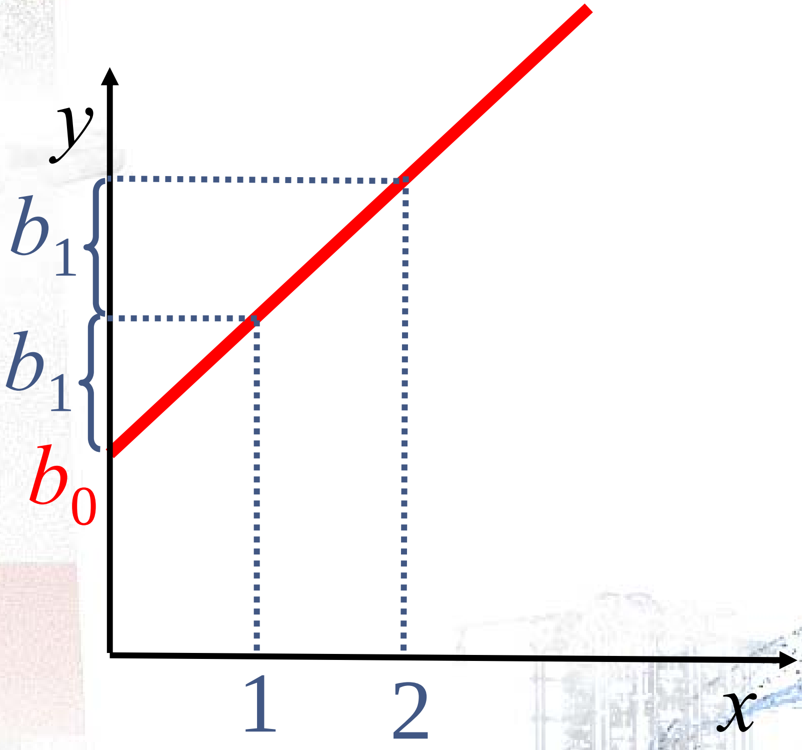
- Example
  - A plumber charges $20 for driving to your house, plus $40 for each hour of work at your home
  - Let
    - $y$ = total charge
    - $x$ = number of hours of work at your house
  - The relationship between $y$ and $x$ is
    - $y$=20 + 40$x$

$y$ intercept

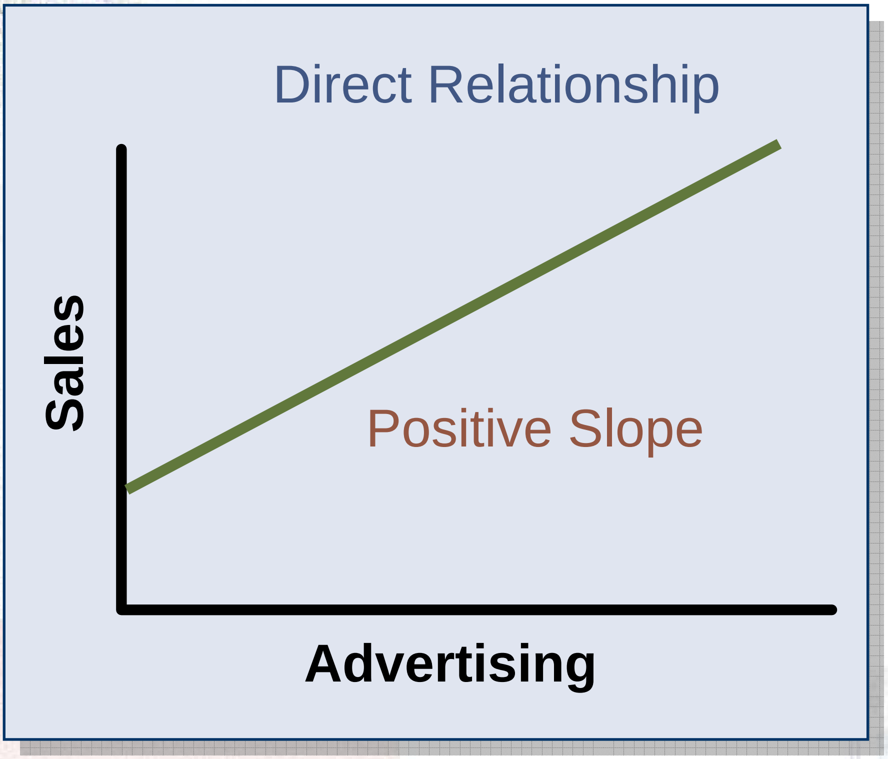
Slope

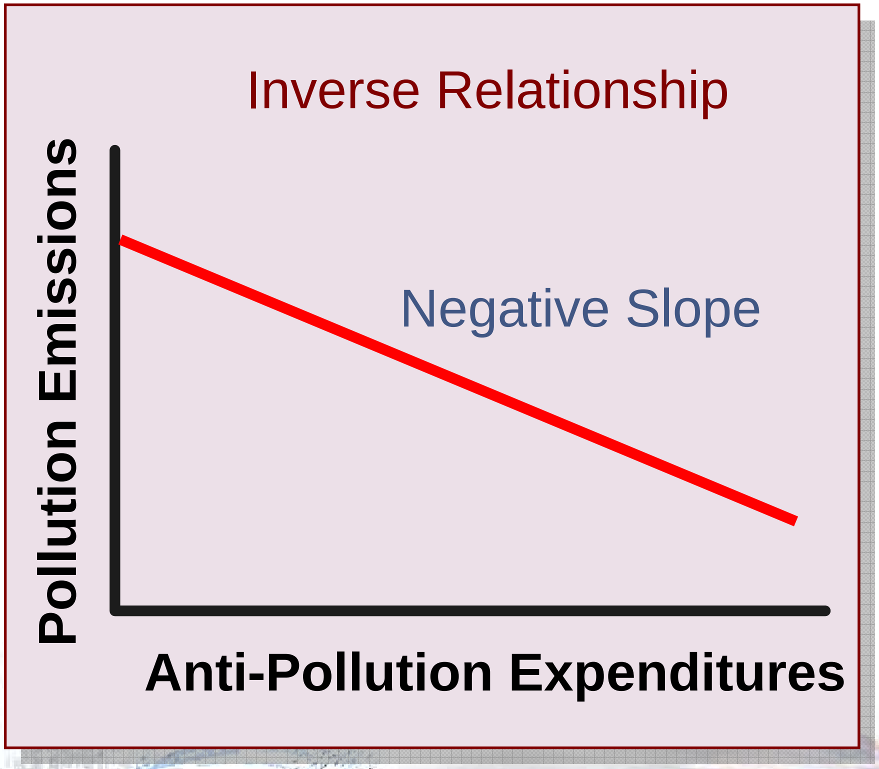
# General Linear Relationship

$$y = b_0 + b_1 x$$



# Direct vs. Inverse Relationship



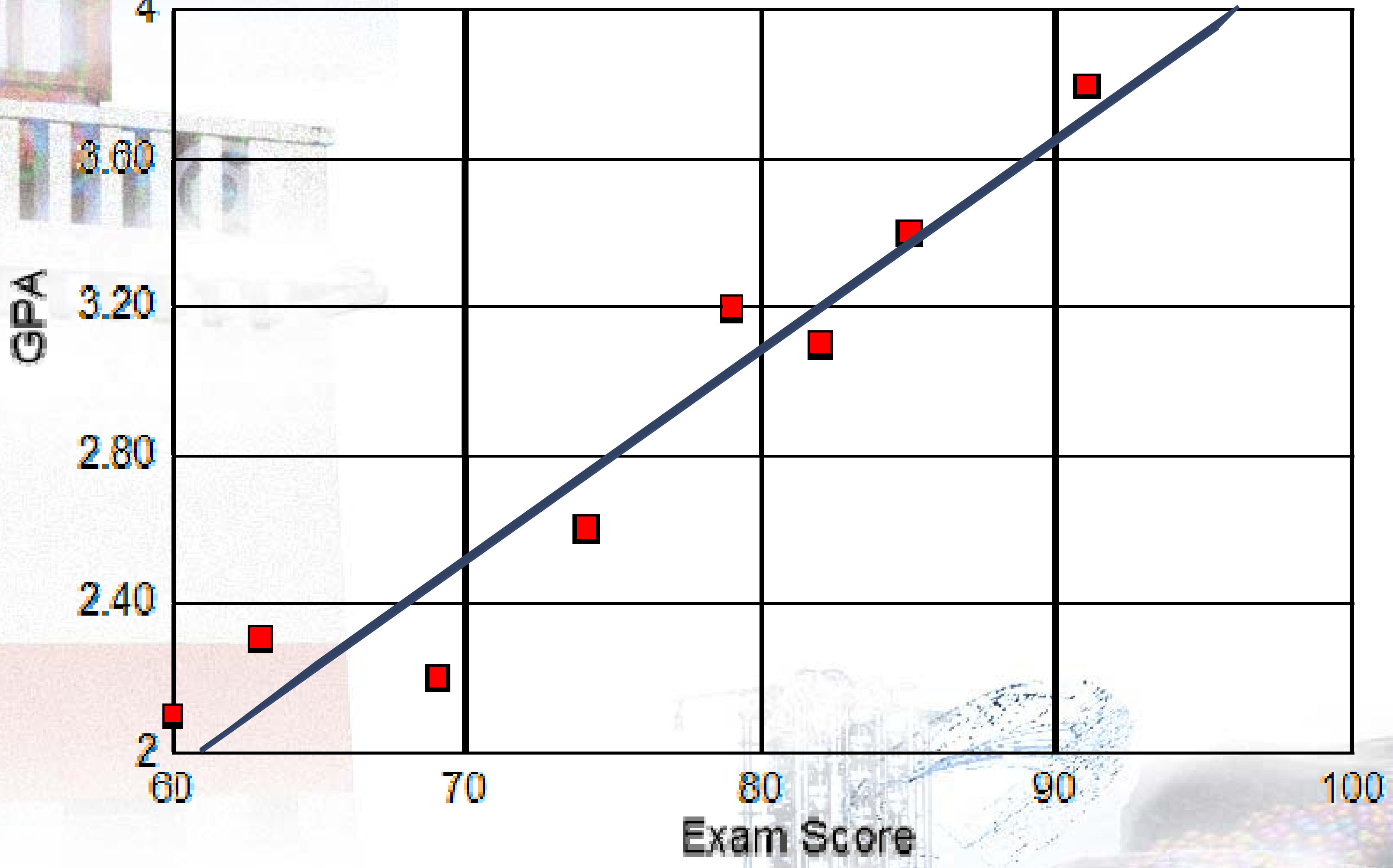


# Parameter Estimation Example

| Student | Exam Score | G.P.A. |
|---|---|---|
| A | 74 | 2.6 |
| B | 69 | 2.2 |
| C | 85 | 3.4 |
| D | 63 | 2.3 |
| E | 82 | 3.1 |
| F | 60 | 2.1 |
| G | 79 | 3.2 |
| H | 91 | 3.8 |

Suppose that your I101 instructor wishes to determine whether any relationship exists between a student's score on an entrance examination and that student's cumulative GPA. A sample of eight students is taken.

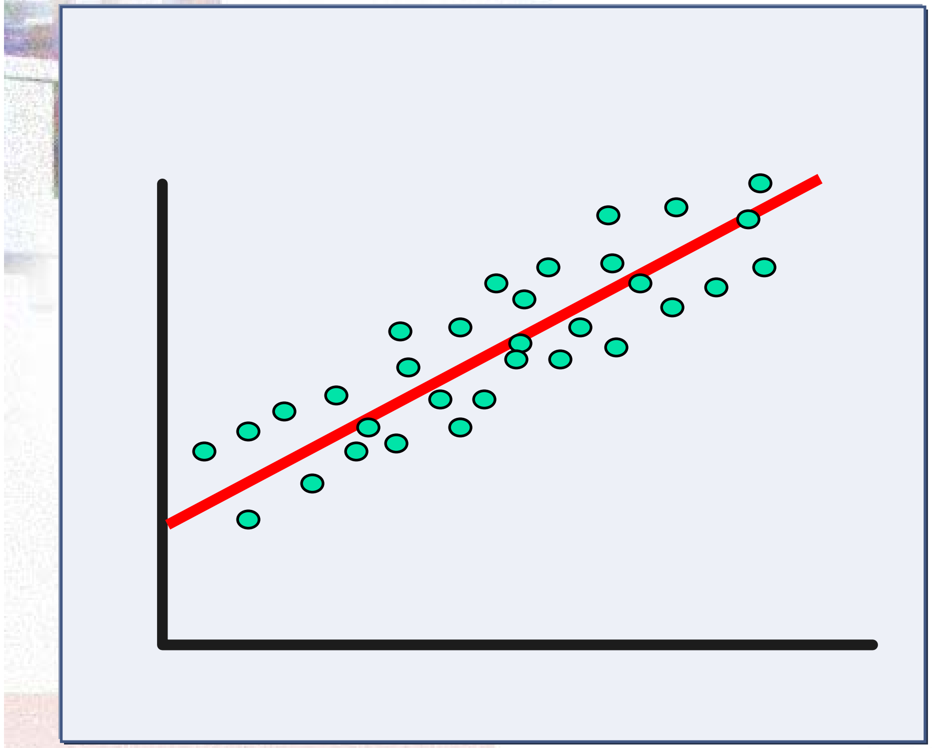
# Scatter Diagram
# GPA vs. Exam Score



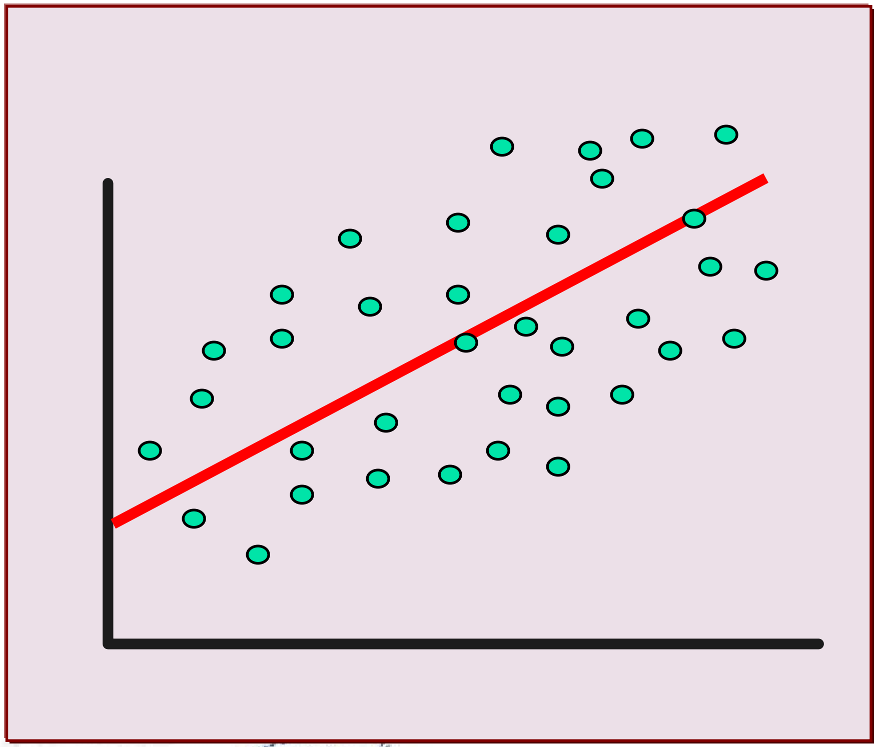
Luis M.Rocha and Santiago Schnell

# Scatter Around Linear Relationship

**More Accurate Estimator of *X, Y* Relationship**



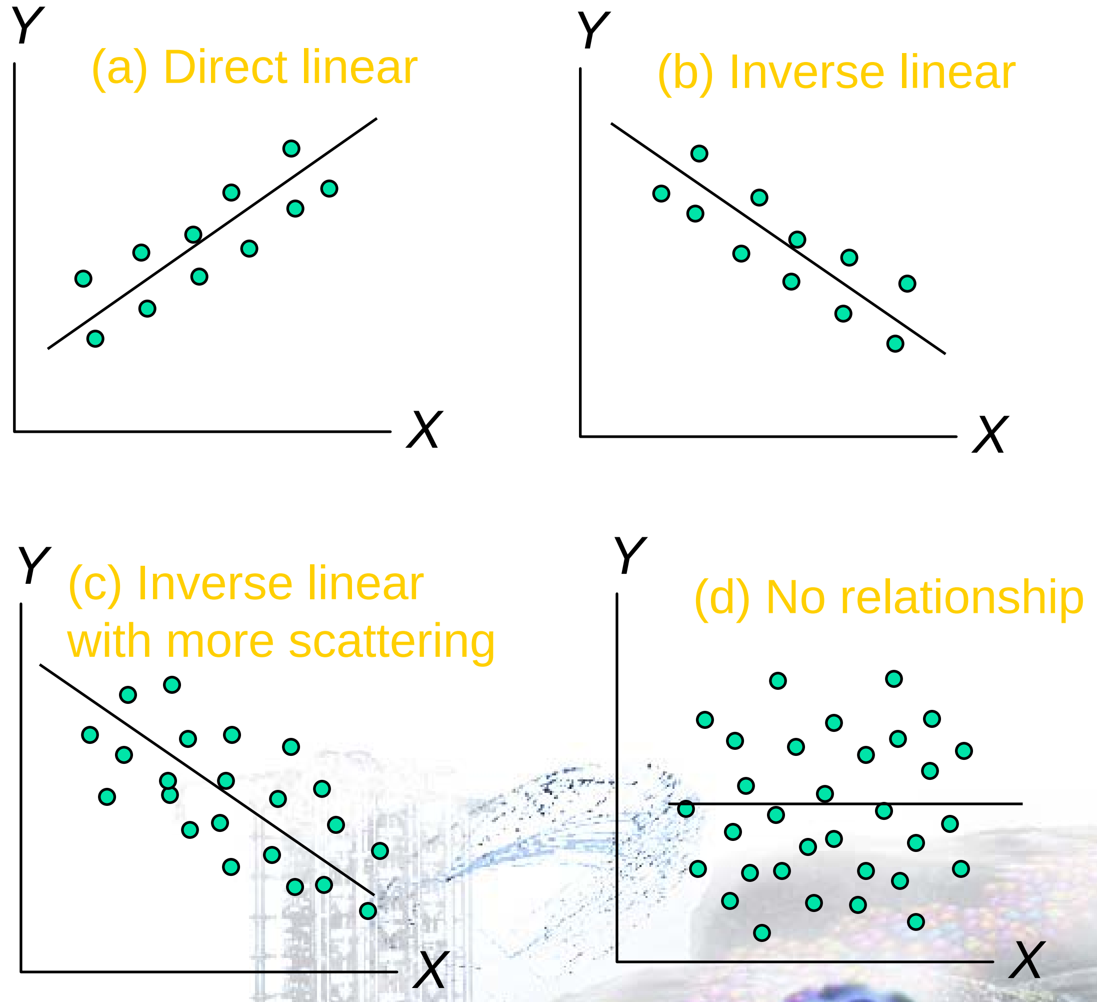
**Less Accurate Estimator of *X, Y* Relationship**



- Larger Error
- More uncertain about inference

Luis M.Rocha and Santiago Schnell

# Possible Relationships Between X and Y in Scatter Diagrams

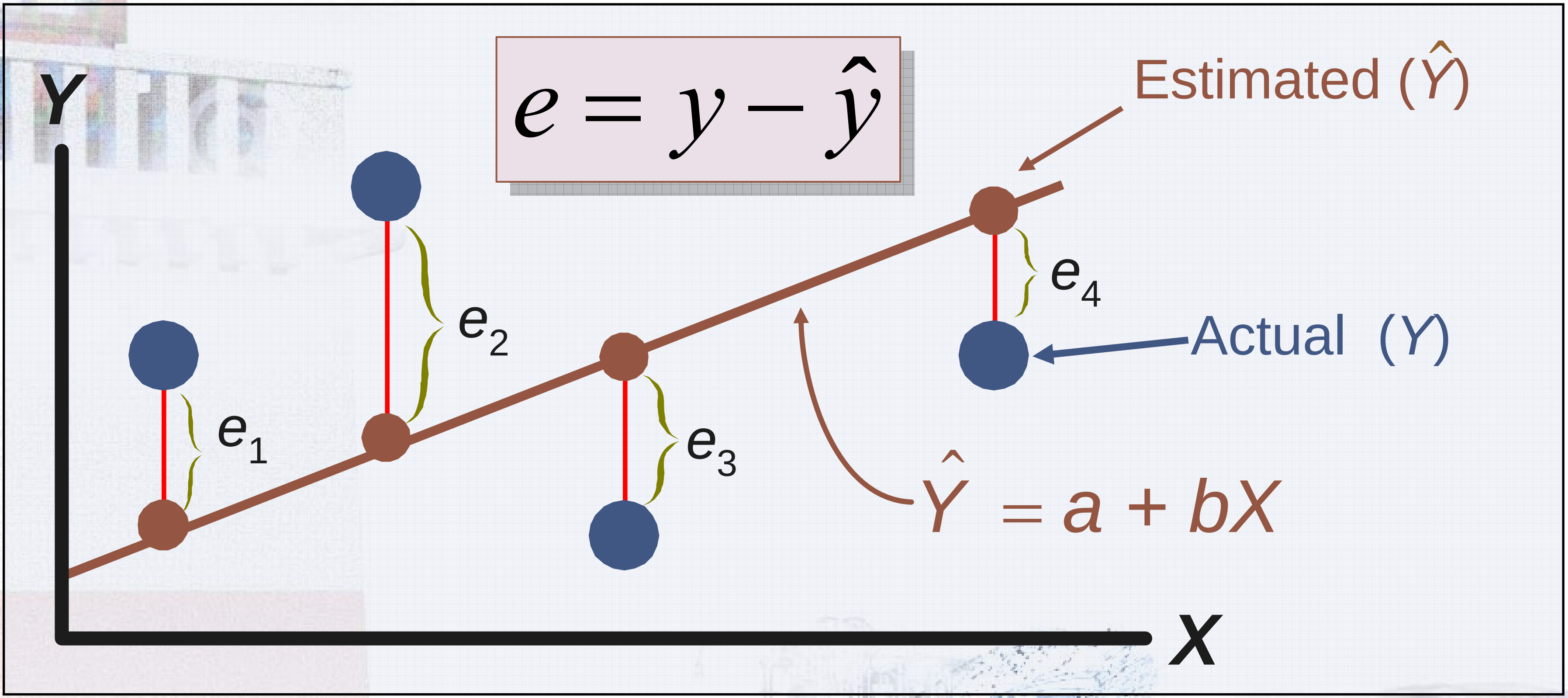


# Comparing data to linear model

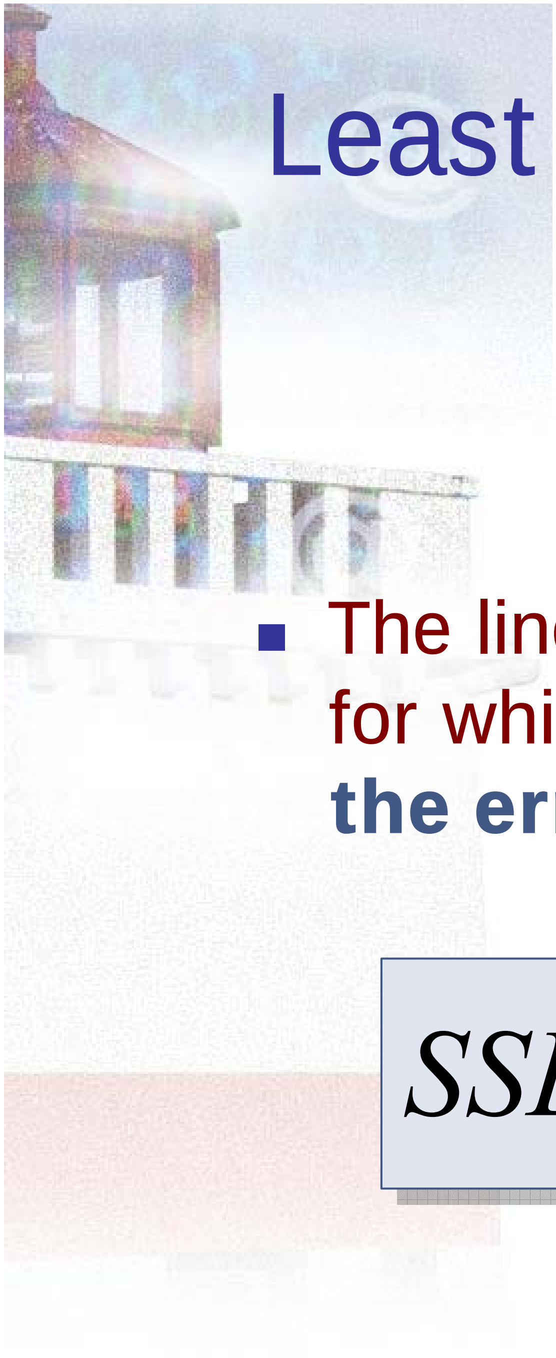
- Accounts differences between actual values ($y$) and estimated or predicted values ($\hat{y}$)
  - Error or residual for a given value

$$\boxed{e = y - \hat{y}}$$

# Errors or Residuals Graphically



$$e = y - \hat{y}$$

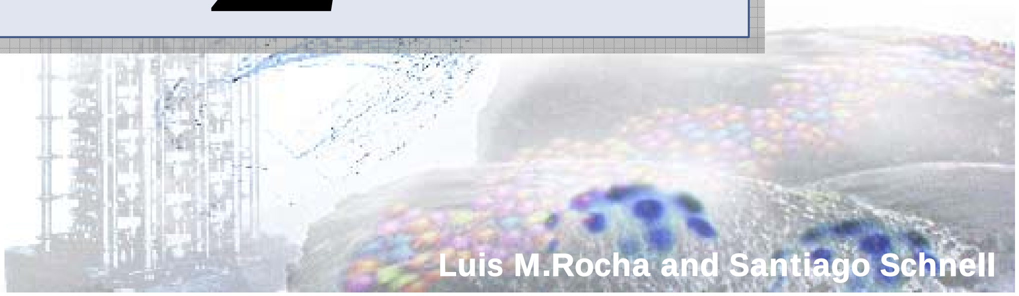


# Least Squares Criterion

$$e = y - \hat{y}$$

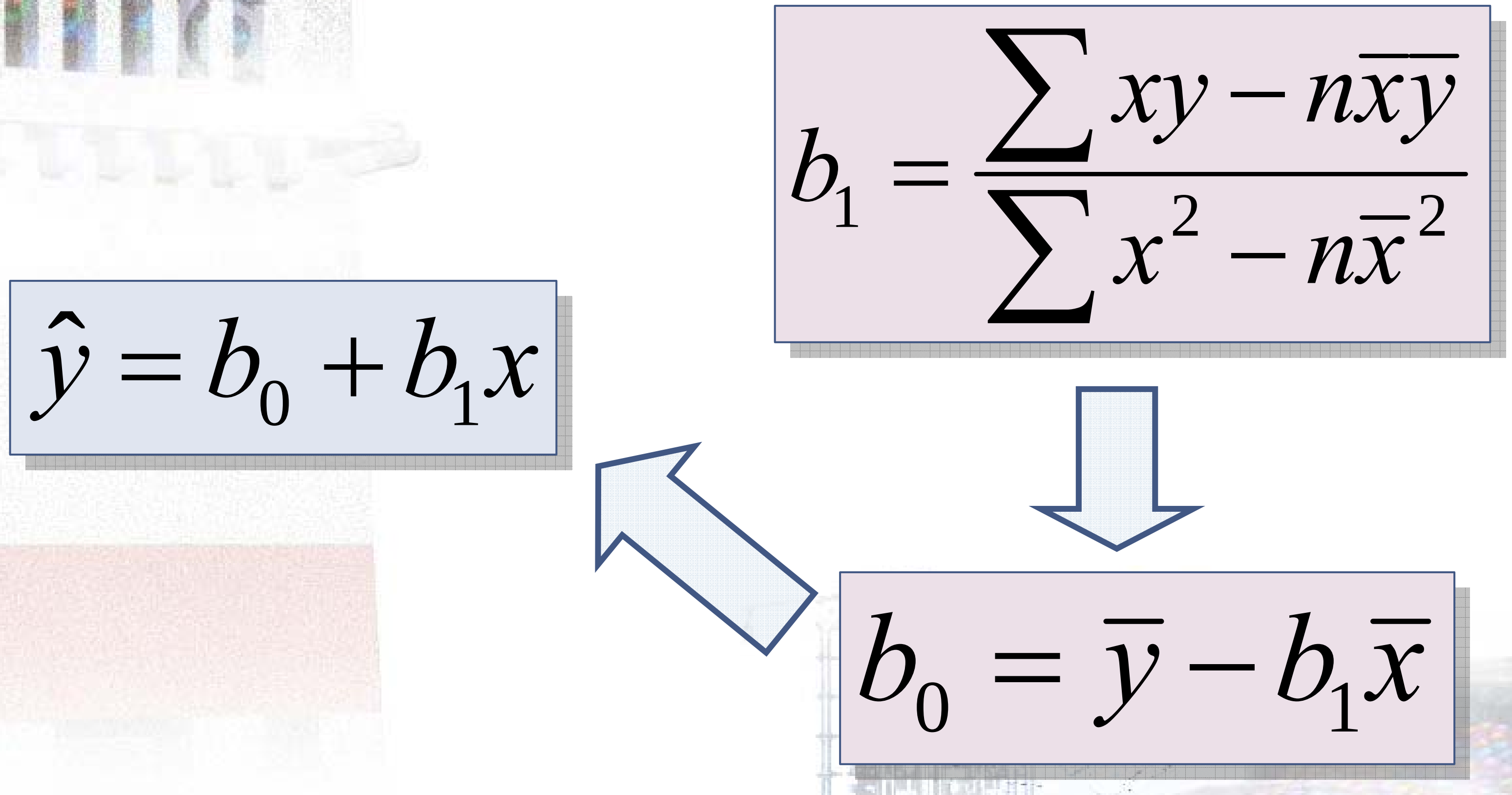
- The line that best fits the data is the one for which the **sum of the squares of the errors** **(SSE)** is smallest

$$SSE = \sum e^2 = \sum \left(y - \hat{y}\right)^2$$



# Method for Regression

- Line of ***best fit*** or ***regression*** based on Least Squares Criterion



$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$\hat{y} = b_0 + b_1 x$$

# Parameter Estimation Example

| Student | Exam ($x$) | GPA ($y$) | $xy$ | $x^2$ |
|---|---|---|---|---|
| A | 74 | 2.6 | 192.4 | 5476 |
| B | 69 | 2.2 | 151.8 | 4761 |
| C | 85 | 3.4 | 289 | 7225 |
| D | 63 | 2.3 | 144.9 | 3969 |
| E | 82 | 3.1 | 254.2 | 6724 |
| F | 60 | 2.1 | 126 | 3600 |
| G | 79 | 3.2 | 252.8 | 6241 |
| H | 91 | 3.8 | 345.8 | 8281 |
| | | | -------- | ------- |
| | | | **1756.9** | **46277** |

$n = 8$

$\bar{x}$=75.375

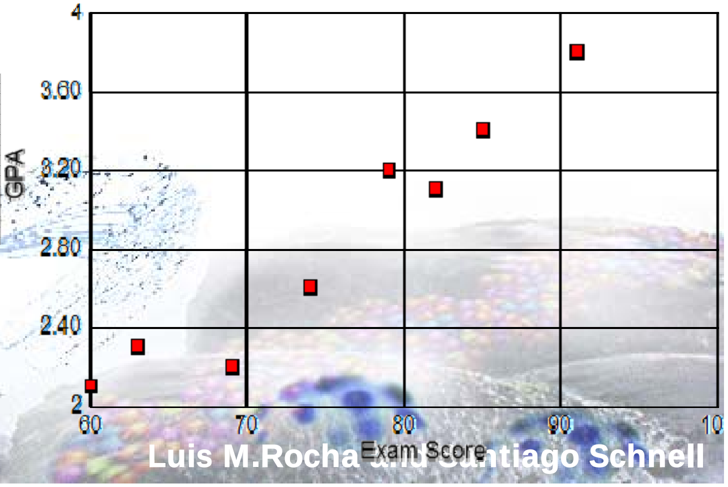
$\bar{y}$=2.8375

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

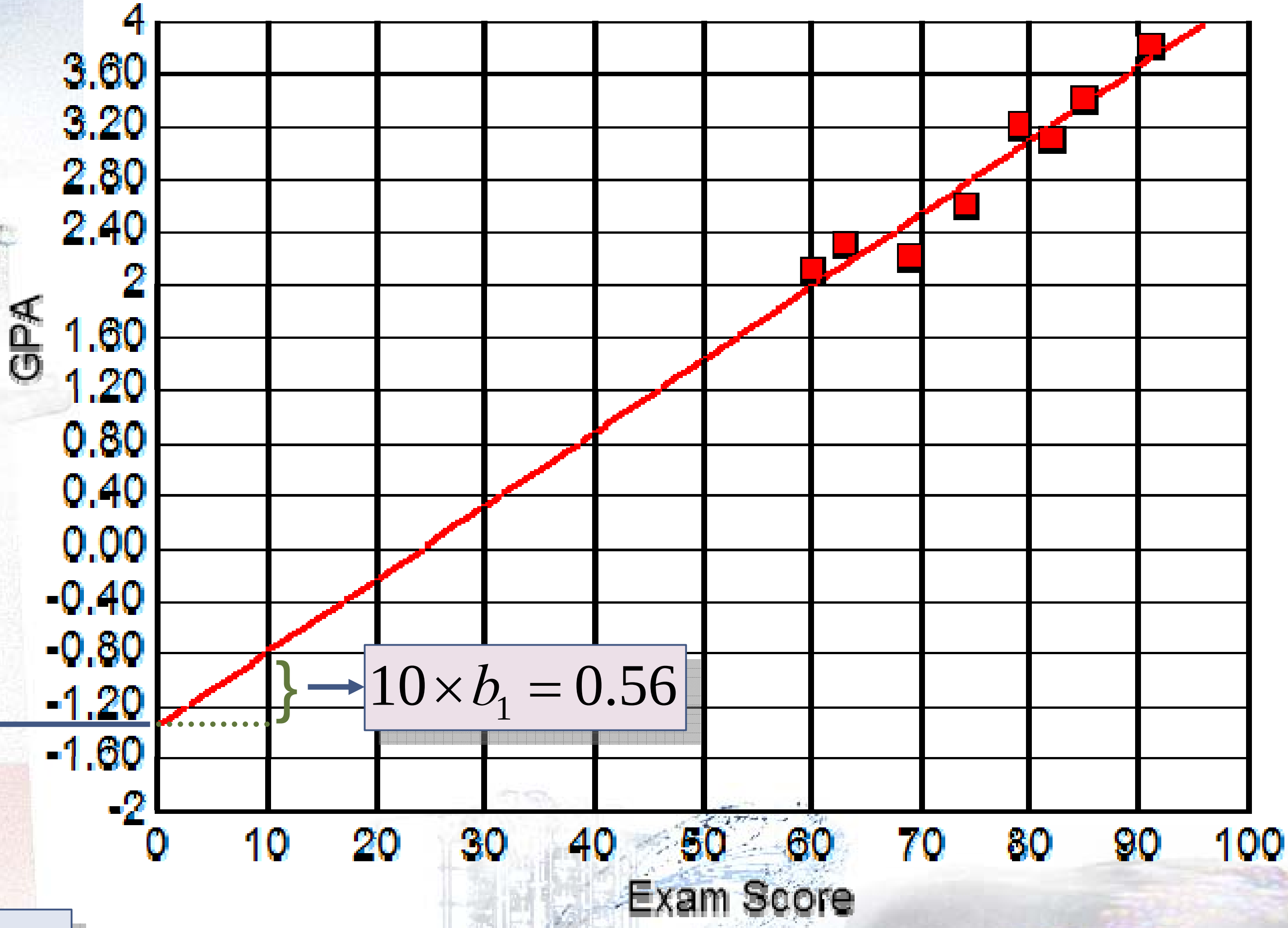
$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{1756.9 - 8 \times 75.375 \times 2.8375}{46277 - 8 \times 75.375^2} = 0.05556$$

$$b_0 = 2.8375 - 0.05556 \times 75.375 = \text{-}1.351$$



# Example Linear Fit



$$10 \times b_1 = 0.56$$

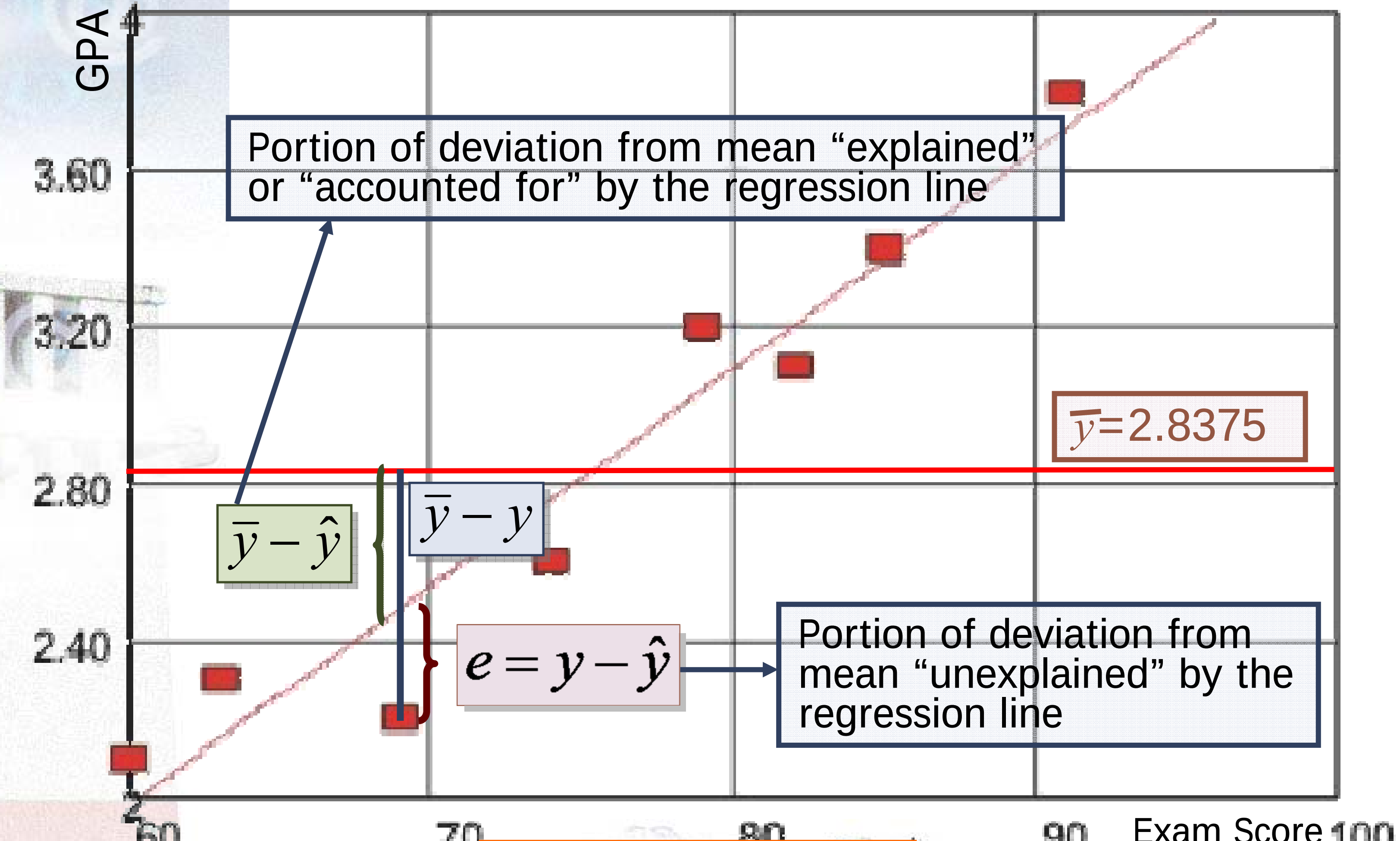
$$b_0 = \text{-}1.351$$

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = 0.05556$$

Luis M.Rocha and Santiago Schnell

# How good is the Linear Fit?



Portion of deviation from mean “explained” or “accounted for” by the regression line

$\bar{y}$=2.8375

$\bar{y}-\hat{y}$

$\bar{y}-y$

$e=y-\hat{y}$

Portion of deviation from mean “unexplained” by the regression line

$TSS = SSE + SSR$

Total Sum of Squares

$$TSS=\sum\left(\bar{y}-y\right)^2$$

Sum of Squares for error

$$SSE=\sum\left(y-\hat{y}\right)^2$$

Sum of Squares for regression

$$SSR=\sum\left(\bar{y}-\hat{y}\right)^2$$

# Coefficient of determination

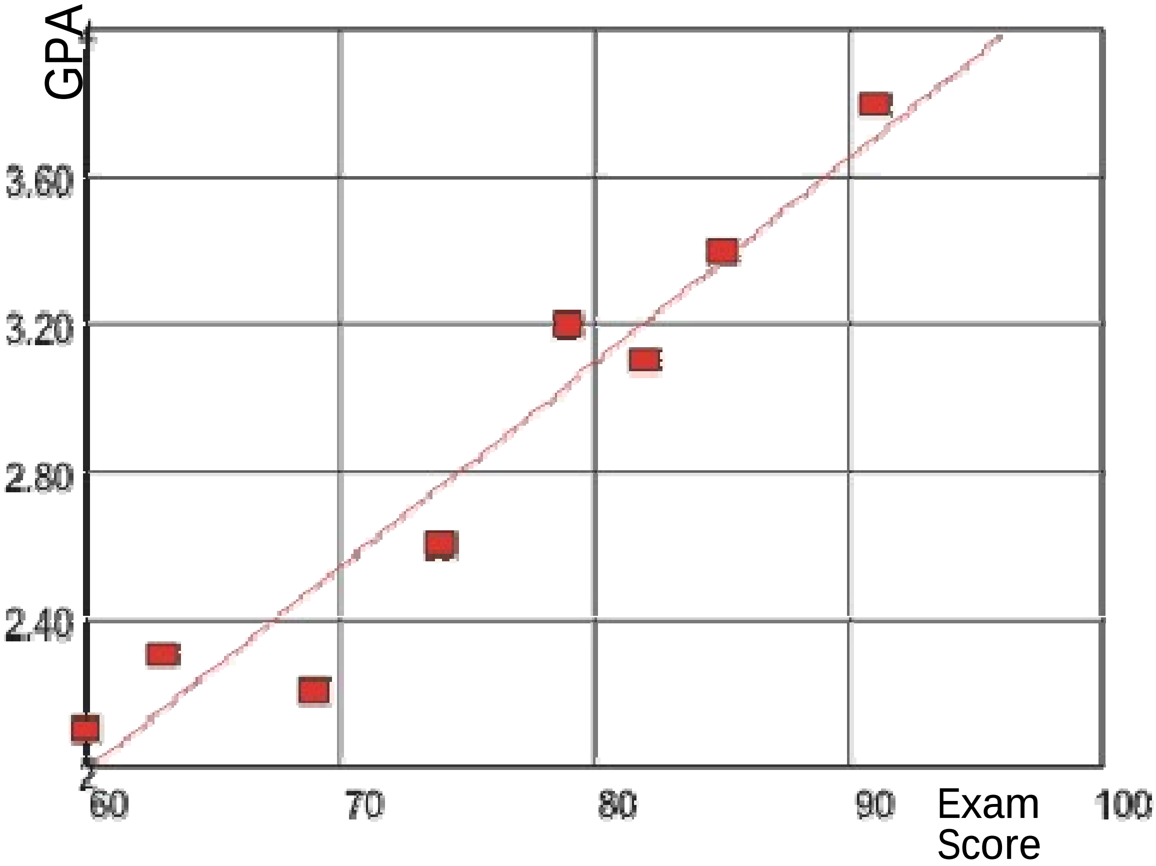
$$\boxed{r^2}$$

- Degree of linear relationship
  - To make a judgment about whether a linear relationship really exists between *x* and *y*.
  - The *proportion* of the variability in y values that is accounted for or *explained by* a linear relationship with x.

$$r^2 = \frac{SSR}{TSS} = \frac{\sum\left(\bar{y} - \hat{y}\right)^2}{\sum\left(\bar{y} - y\right)^2}$$

# Example: Coefficient of determination

$$R^2 = 0.93$$



93% of the GPA variability is explained by the Exam Score (with a linear relationship)

# Coefficient of Correlation

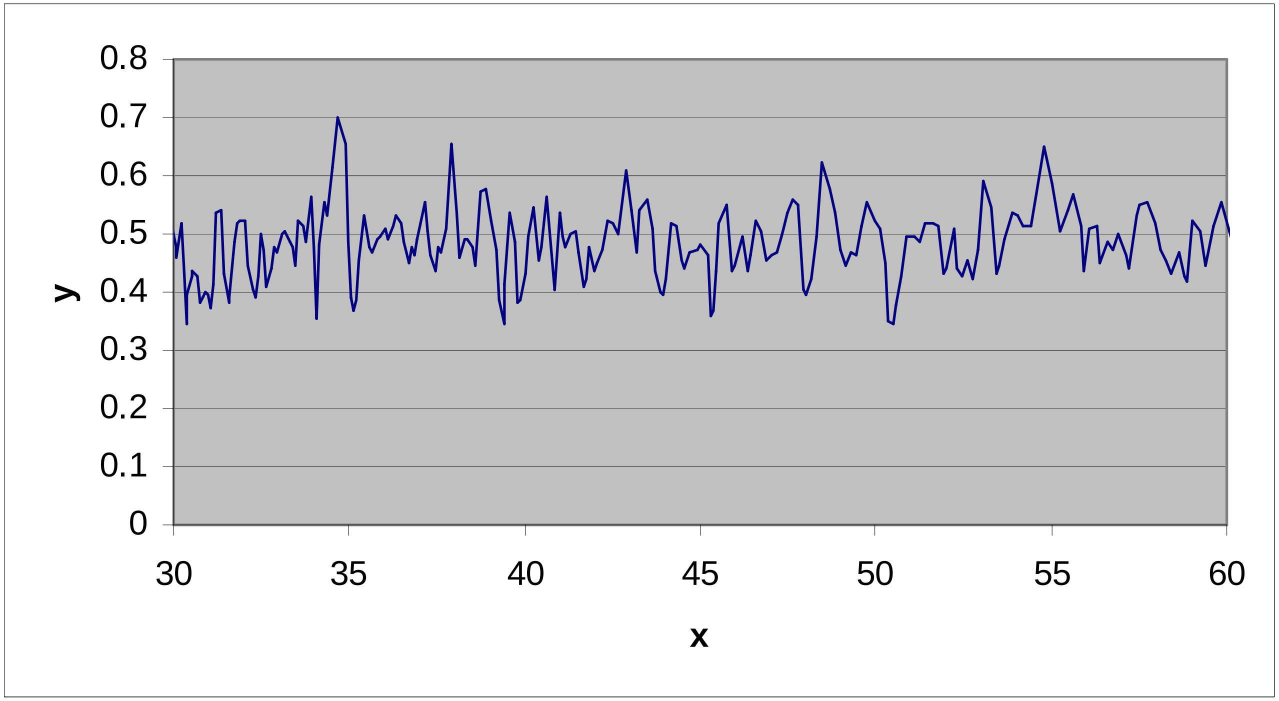
- Degree of linear relationship
  - $r^2$ is easier to interpret

$$\boxed{r}$$

- Allows us to infer how good a linear model is
  - The quality of our inferences: our degree of ***uncertainty***

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y} = \frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right)-\left(\sum x\right)^2}\,.\sqrt{n\left(\sum y^2\right)-\left(\sum y\right)^2}}$$

# ... however, sometimes we cannot fit data to a straight line!



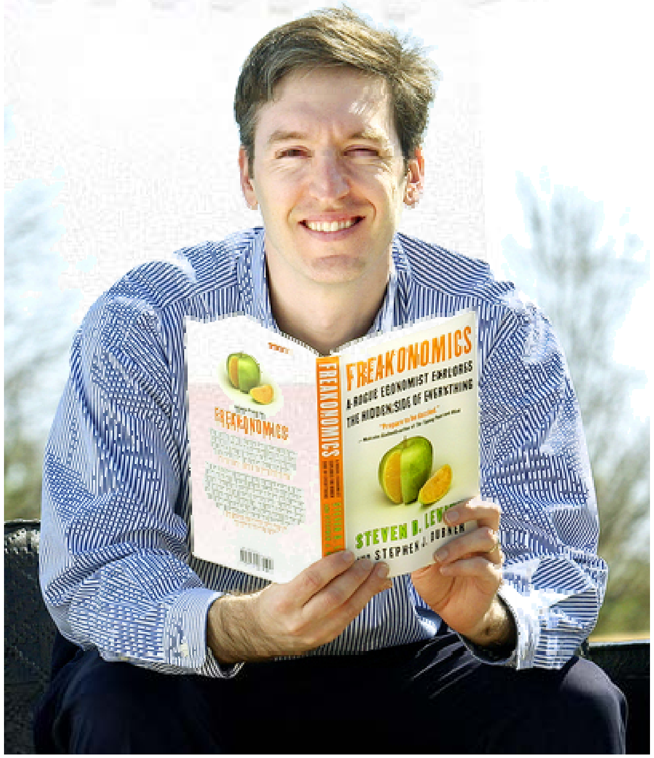
Why cannot we model some processes with lines?

- Large measurement errors
- Presence of noise
- The process is random

FREAKONOMICS A ROGUE ECONOMIST EXPLORES THE HIDDEN SIDE OF EVERYTHING

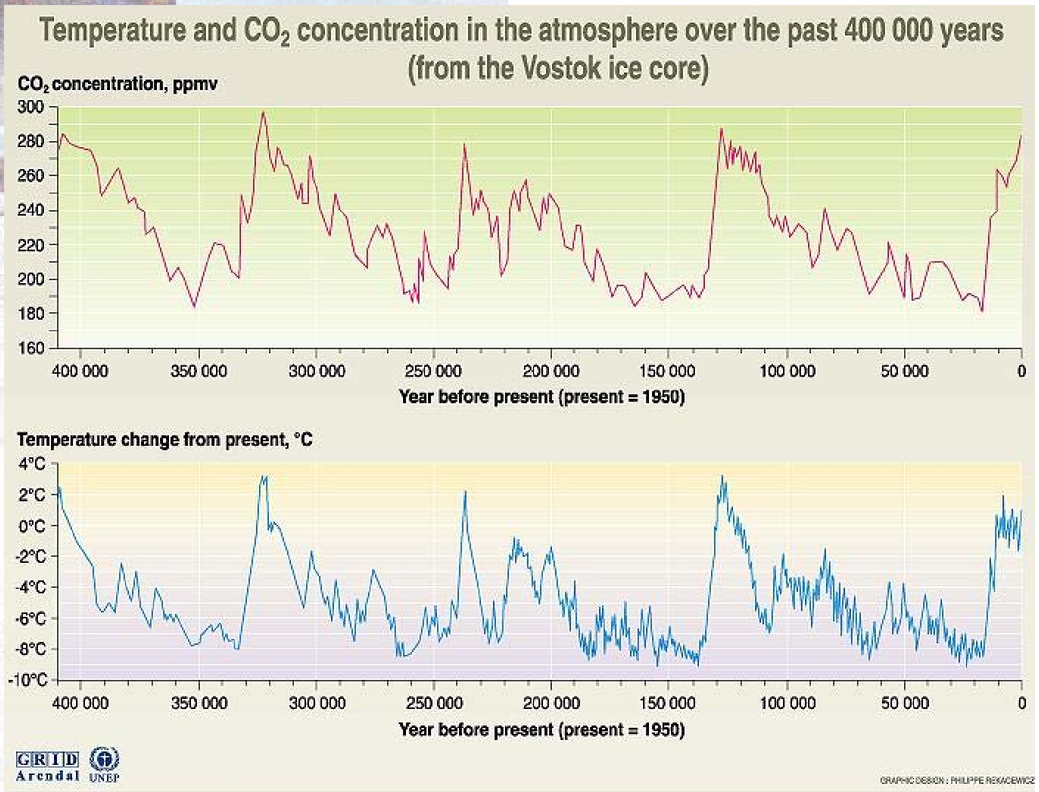
BY STEVEN D. LEVITT
AND STEPHEN J. DUBNER

Challenge: correlation does not prove causation!

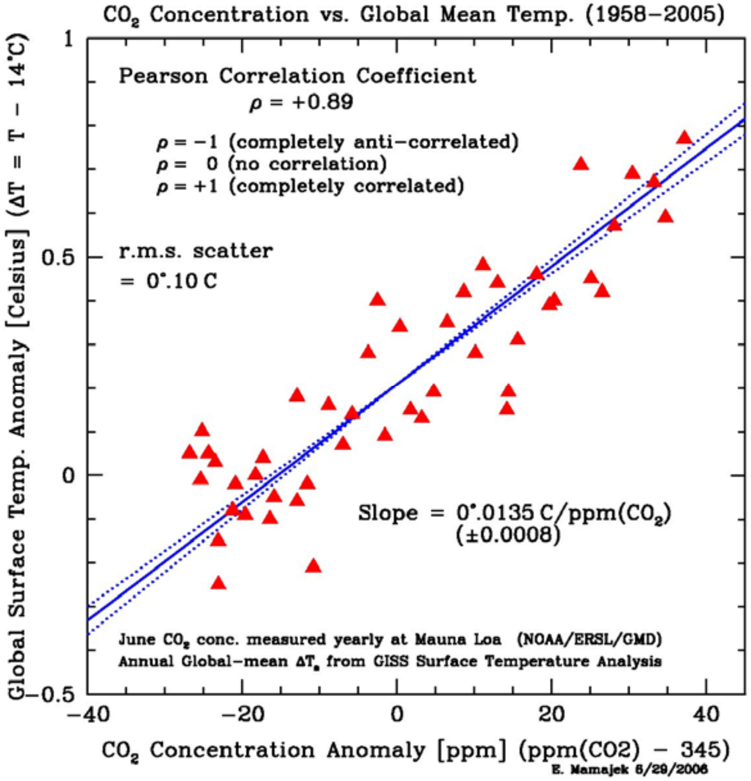


- Cheating of School Teachers and Sumo Wrestlers
- The incentives of real estate agents
- Why do drug dealers still live with their moms?
- Parenthood, names, and social status?
- Row vs. Wade and Low Crime Rates

# Global Warming



Source: J.R. Petit, J. Jouzel, et al. Climate and atmospheric history of the past 420 000 years from the Vostok ice core in Antarctica, Nature 399 (3JUne), pp 429-436, 1999.

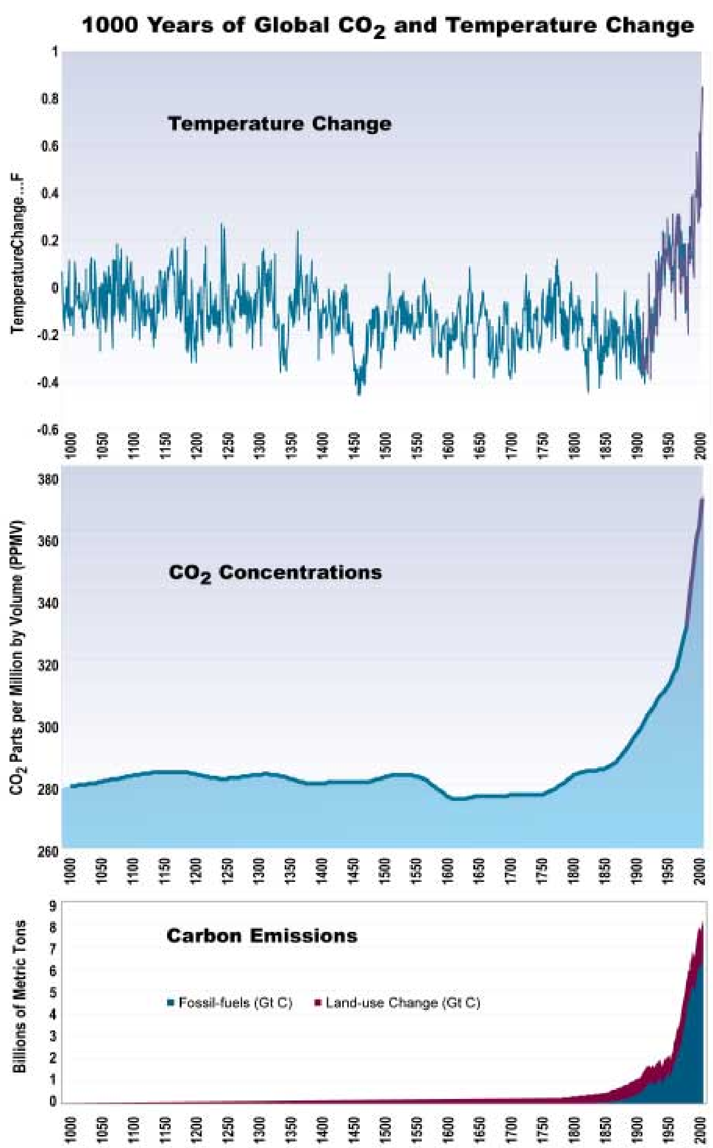


Erik Mamajek:

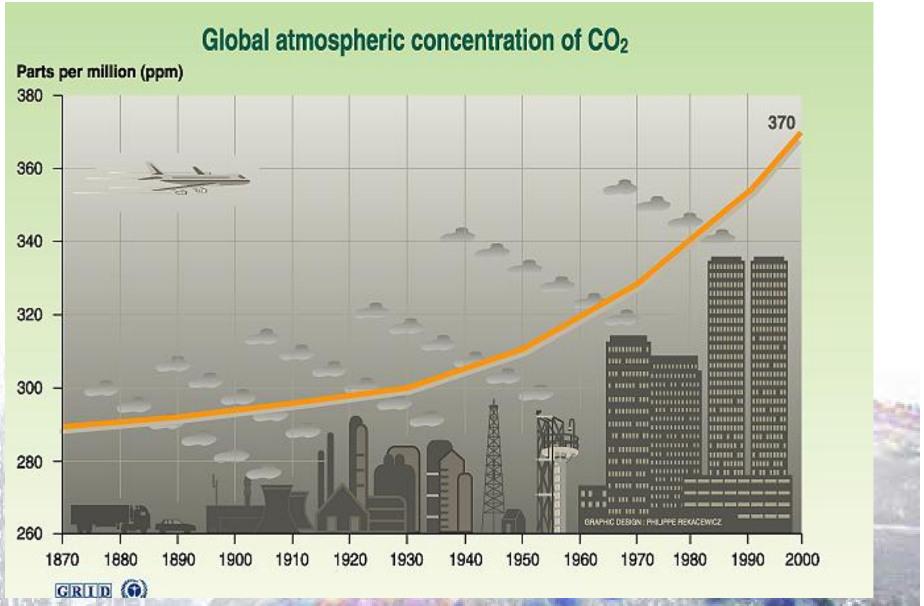
http://cfa-www.harvard.edu/~emamajek/

http://www.grida.no/climate/vital

# Global Warming



1000 Years of Global $CO_2$ and Temperature Change



Global atmospheric concentration of $CO_2$

# Frequency Analysis and Cryptography

- Cryptography
  - Derived from the Greek word *Kryptos*: hidden
- See Simon Singh's The Code Book CD-ROM
  - The Vigenère Code

# Next Class!

- Topics
  - More Inductive Reasoning Modeling
    - Probability and Uncertainty
- Readings for Next week
  - @ *infoport*
  - From course package
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 109-134)
      - OPTIONAL: Chapter 4 (pages 135-140)
      - Chapter 13 (pages 151-159)
      - Chapter 5  (pages 141-144)
- Lab 8
  - Intro to Statistical Analysis using Excel