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!"
Readings until now

- Lecture notes
  - Posted online
    - [http://informatics.indiana.edu/rocha/i101](http://informatics.indiana.edu/rocha/i101)
      - *The Nature of Information*
      - *Technology*
      - *Modeling the World*
  - [http://infoport.blogspot.com](http://infoport.blogspot.com)

- From course package
    - 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)
    - Chapter 2: Classical Logic (pp. 87-97)
    - Chapter 3: Classical Set Theory (pp. 98-103)
    - 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

- 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
Linear Relationship

**Example**

- A plumber charges $20 for driving to your house, plus $40 for each hour of work at your home

- Let
  - $y = \text{total charge}$
  - $x = \text{number of hours of work at your house}$

- The relationship between $y$ and $x$ is
  - $y = 20 + 40x$
General Linear Relationship

\[ y = b_0 + b_1 x \]
Direct vs. Inverse Relationship

- Direct Relationship
  - Positive Slope
  - Sales vs. Advertising

- Inverse Relationship
  - Negative Slope
  - Pollution Emissions vs. Anti-Pollution Expenditures
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.

<table>
<thead>
<tr>
<th>Student</th>
<th>Exam Score</th>
<th>G.P.A.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>74</td>
<td>2.6</td>
</tr>
<tr>
<td>B</td>
<td>69</td>
<td>2.2</td>
</tr>
<tr>
<td>C</td>
<td>85</td>
<td>3.4</td>
</tr>
<tr>
<td>D</td>
<td>63</td>
<td>2.3</td>
</tr>
<tr>
<td>E</td>
<td>82</td>
<td>3.1</td>
</tr>
<tr>
<td>F</td>
<td>60</td>
<td>2.1</td>
</tr>
<tr>
<td>G</td>
<td>79</td>
<td>3.2</td>
</tr>
<tr>
<td>H</td>
<td>91</td>
<td>3.8</td>
</tr>
</tbody>
</table>
Scatter Around Linear Relationship

More Accurate Estimator of $X, Y$ Relationship

Less Accurate Estimator of $X, Y$ Relationship

- Larger Error
- More uncertain about inference
Possible Relationships Between X and Y in Scatter Diagrams

(a) Direct linear

(b) Inverse linear

(c) Inverse linear with more scattering

(d) No relationship
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

$$e = y - \hat{y}$$
Errors or Residuals Graphically

\[
e = y - \hat{y}
\]

\[
\hat{Y} = a + bX
\]
Least Squares Criterion

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

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

$$SSE = \sum e^2 = \sum (y - \hat{y})^2$$
Method for Regression

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

\[ \hat{y} = b_0 + b_1 x \]

\[ b_1 = \frac{\sum (xy - nx\bar{y})}{\sum (x^2 - n\bar{x}^2)} \]

\[ b_0 = \bar{y} - b_1 \bar{x} \]
### Parameter Estimation Example

#### Data Table

<table>
<thead>
<tr>
<th>Student</th>
<th>Exam ($x$)</th>
<th>GPA ($y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>74</td>
<td>2.6</td>
</tr>
<tr>
<td>B</td>
<td>69</td>
<td>2.2</td>
</tr>
<tr>
<td>C</td>
<td>85</td>
<td>3.4</td>
</tr>
<tr>
<td>D</td>
<td>63</td>
<td>2.3</td>
</tr>
<tr>
<td>E</td>
<td>82</td>
<td>3.1</td>
</tr>
<tr>
<td>F</td>
<td>60</td>
<td>2.1</td>
</tr>
<tr>
<td>G</td>
<td>79</td>
<td>3.2</td>
</tr>
<tr>
<td>H</td>
<td>91</td>
<td>3.8</td>
</tr>
</tbody>
</table>

#### Statistics

- $n = 8$
- $\bar{x} = 75.375$
- $\bar{y} = 2.8375$
- $\sum xy = 1756.9$
- $\sum x^2 = 46277$
- $\sum y = 22.71$
- $\sum x = 600$

#### Calculations

- $b_1 = \frac{\sum xy - n\bar{xy}}{\sum x^2 - n\bar{x}^2} = \frac{1756.9 - 8 \times 75.375 \times 2.8375}{46277 - 8 \times 75.375^2} = 0.05556$
- $b_0 = \bar{y} - b_1 \bar{x} = 2.8375 - 0.05556 \times 75.375 = -1.351$
Example Linear Fit

\[ y = b_0 + b_1 x \]

- \( b_0 = -1.351 \)
- \( b_1 = 0.05556 \)
- \( 10 \times b_1 = 0.56 \)
How good is the Linear Fit?

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

\[
\overline{y} - \hat{y}
\]

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

\[
e = y - \hat{y}
\]

Total Sum of Squares

\[
TSS = \sum (\overline{y} - y)^2
\]

Sum of Squares for error

\[
SSE = \sum (y - \hat{y})^2
\]

Sum of Squares for regression

\[
SSR = \sum (\overline{y} - \hat{y})^2
\]

TSS = SSE + SSR
Coefficient of determination

- 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 \( x \).

\[
r^2 = \frac{SSR}{TSS} = \frac{\sum (\bar{y} - \hat{y})^2}{\sum (\bar{y} - y)^2}
\]
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
- Allows us to infer how good a linear model is
- The quality of our inferences: our degree of *uncertainty*

\[
r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1)s_x s_y} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}}
\]
... 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
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?
- Roe vs. Wade and Low Crime Rates
Global Warming

Temperature and CO₂ concentration in the atmosphere over the past 400,000 years (from the Vostok ice core)

CO₂ concentration, ppmv

Temperature change from present, °C

Year before present (present = 1950)


Pearson Correlation Coefficient

\[ \rho = +0.89 \]

\[ \rho = -1 \] (completely anti-correlated)

\[ \rho = 0 \] (no correlation)

\[ \rho = +1 \] (completely correlated)

r.m.s. scatter

\[ = 0.10°C \]

Slope \[ = 0.0135°C/\text{ppm}(CO₂) \]

\[ \pm 0.0008 \]

June CO₂ conc. measured yearly at Mauna Loa (NOAA/ERSL/GMD)

Annual Global-mean \( \Delta T \) from GISS Surface Temperature Analysis

Erik Mamajek:

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

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

1000 Years of Global CO₂ and Temperature Change

Temperature Change

CO₂ Concentrations

Carbon Emissions

Global atmospheric concentration of CO₂
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
    - 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