



Statistical Modeling as an Introductory Course

Daniel Kaplan
Macalester College



Designing an Introductory Statistics Course from Scratch

The audience:

Natural and social science students
sympathetic to measuring things: e.g.,
majors in biology, economics and business,
psychology

These students take follow-up course or
labwork:

Economics: econometrics

Biology: multi-way ANOVA

Psychology: repeated measures ANOVA



Students should develop skills in:

- Descriptive statistics and graphics
- Bias, variability, habits of collecting data
 - e.g., randomization and control
- Probability
 - ~~Pairs of dice, sets of coins, and so on~~
- Ideas of statistical inference
 - ~~t-test, p, z-tests, simple regression, ANOVA~~
- Understanding relations among variables
- Conditioning
 - Conditional probabilities, Bayesian thought



Why Not t-,p-,z-tests?

- A formal way to a simple question:
 - Are these two groups different?
- Technical apparatus is a black box:
 - Students focus on the output: a p-value
- Sends a message that statistics is about formalism, not thinking. A hoop to jump through.
- Teaches students to ignore covariates or mystifies them: Simpson's *paradox*, *not* "situation"
- Rote methods: pick the right test and do it.



Modeling instead of t-p-z

Models: linear, glm, logistic,...

- Emphasize the various sorts of relationships among variables
- Multivariate: takes into account confounding or contributing variables
- When you finish a stats course, Simpson's shouldn't be a paradox.

Example: Air bags and safety from the last issue of Chance. Lower accident fatality rate in cars with airbags, but "if people in cars are more likely to be wearing their seatbelts, then perhaps some of the apparent effectiveness of airbags is really due to increased seatbelt use. ...[O]ccupants with airbags are much more likely to be wearing seatbelts properly."¹

¹ M.C. Meyer and T. Finney, "Who Wants Airbags?" Chance 18(2), 3:15 (2005)



Why Multivariate Modeling?

- **Applicability:** Client fields use these techniques
 - Economics: multiple regression
 - Psychology: repeated measures
 - Biology: ANOVAShouldn't people who understand the mathematical structure be involved in teaching the "advanced" techniques?
- **Quantification:** Create the habit of looking at the strength of a relationship, not just the significance.
- **Creativity:** Make introductory statistics give techniques that let students express their own interests.
- **Empowerment:** Move from "paradox" to understanding.



How to teach multivariate modeling to intro. students?

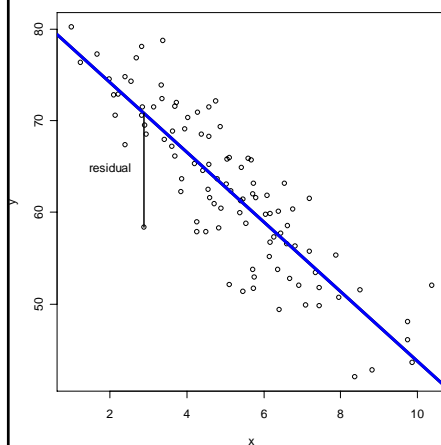
Replace algebra with **computation**, **simulation**, and **geometry**.

- Simulation
 - Confidence intervals via bootstrapping
 - Hypothesis testing via randomization of explanatory variables
- Geometry
 - Regression as projection
 - ANOVA as pythagorean vector decomposition.
 - p-values from subtended angles

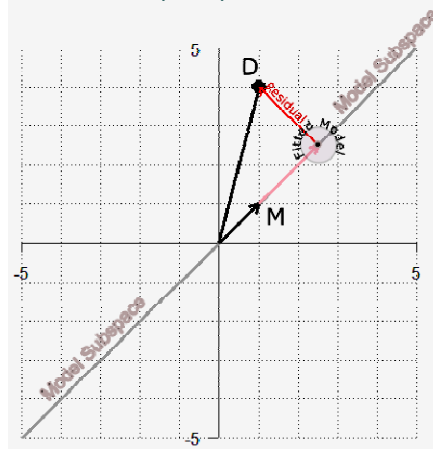


Regression

Scatter-plot presentation



Vector-space presentation



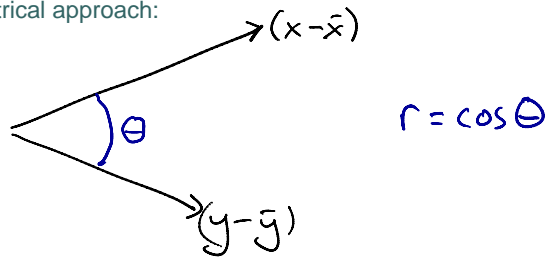


Correlation

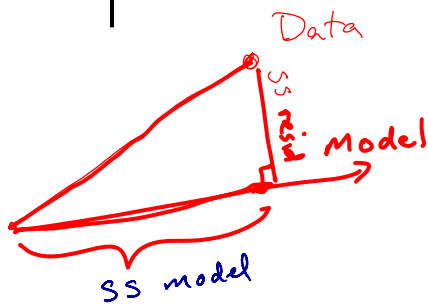
Algebraic approach:

$$r = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{(\sum_{i=1}^n (x_i - \bar{x})^2) (\sum_{i=1}^n (y_i - \bar{y})^2)}}$$

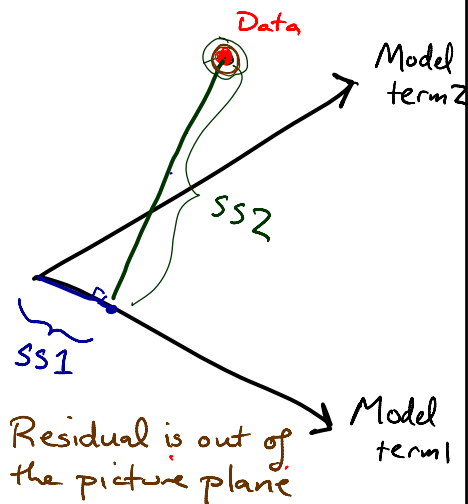
Geometrical approach:



Analysis of Variance



Sums of squares are just the length-squared of projection and residual vectors.



● ● ● | Hypothesis Testing

Is the mean of two measurements, 1 & 4, significantly different from 0?

Conventional:

$$m = 2.5, \quad n = 2$$

$$s = 2.12$$

$$t = m / (s / \sqrt{n}) = 1.666$$

df = 1 use table to find

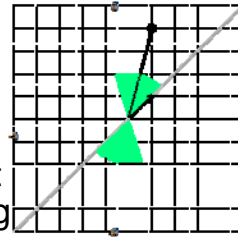
$$p = 0.3440$$

Geometrical:

Find angle between (1,4) and (1,1).

30.9 degrees

What is the prob. that a randomly pointing vector is closer to (1,4) than (1,1)?



● ● ● | It shouldn't be hard...



Is the mean of (1,2,3) significantly different from zero?

- Angle from (1,1,1) is 49 degrees.
- What fraction of the northern hemisphere is above 41 deg latitude?



Status of the Project

- Teaching class for the 4th year.
- Now the main intro. class (w/Calc pre-req.)
- Initial development supported by Howard Hughes Medical Institute.
- Ongoing support from the Keck Foundation for
 - dissemination
 - forging stronger curricular links with economics, sociology, psychology, geography.
 - development of text materials
- Builds on a calculus project that brings mathematics of multiple variables (vectors, subspaces, gradients, partial derivs., multivariate polynomials) to traditional Calc I students.



Going further ...

- We're looking for collaborators to help us develop and test approaches suitable for a wide range of students at diverse sorts of institutions.
- Notes and "labs" available now.
- Textbook for summer 2007.
- Keck-supported summer workshops planned for 2007 and 2008.
- A 3rd tenure-track statistics faculty position at Macalester is now open. Apply!
- Contact: kaplan@macalester.edu



An in-class activity:
Sampling from the $F(2,2)$
distribution by scattering
small slips of paper.