

A Geometrical Approach to Introductory Statistics

Daniel T. Kaplan
Macalester College

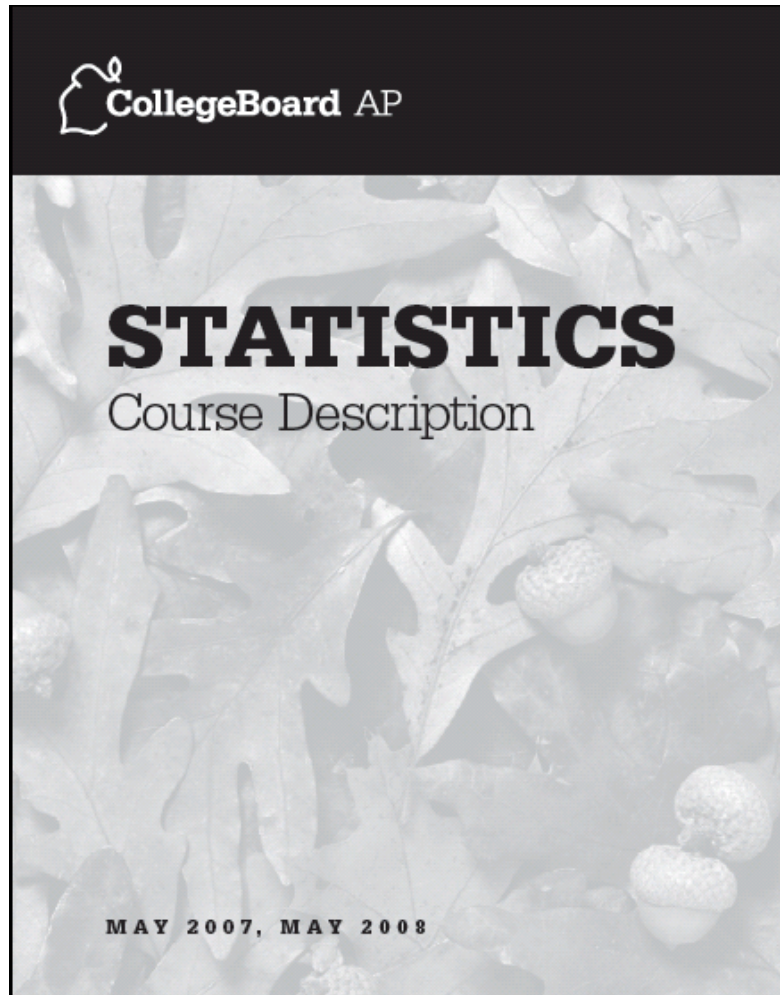
Our Project Goal

Make multivariable statistical ideas accessible to introductory students so they can use them effectively in their own work.

- An introductory course such as the AP course stops at the t-test --- “Are these two groups different?”
 - A college-level course might go a bit further, touching superficially on one-way ANOVA, two-way ANOVA, multiple regression.

Example:

The AP Statistics Curriculum



Tests of significance

1. Logic of significance testing, null and alternative hypotheses; p-values.
2. Large sample test for a proportion
3. Large sample test for a difference between two proportions
4. Test for a mean
5. Test for a difference between two means (unpaired and paired)
6. Chi-square test for goodness of fit, homogeneity of proportions, and independence (one- and two-way tables)
7. Test for the slope of a least-squares regression line

How Do We Move Forward to Multivariable Topics?

- Client departments often offer follow-up, discipline-specific “methods” courses. At Macalester:
 - Econometrics: multiple regression as a starting point
 - Psychology: analysis of variance
 - Ecology and Physiology: Lab exercises using analysis of covariance
- Sometimes these are “turn the crank” approaches that fail to give students a theoretical framework for understanding.

Another Approach to Moving Forward

- Adopt a formalism that makes “advanced” topics elementary and obvious.
- The formalism should be informative, not just a way to turn the crank.

One Formalism: Formulas and Algorithms

I. Descriptive Statistics

$$\bar{x} = \frac{\sum x_i}{n}$$

$$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

Two-Sample

Statistic	Standard Deviation of Statistic
Difference of sample means	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ Special case when $\sigma_1 = \sigma_2$ $\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

These can be useful, but they are not the only way of doing things.

Another Formalism: The Scatter Plot

$$\hat{y} = b_0 + b_1x$$

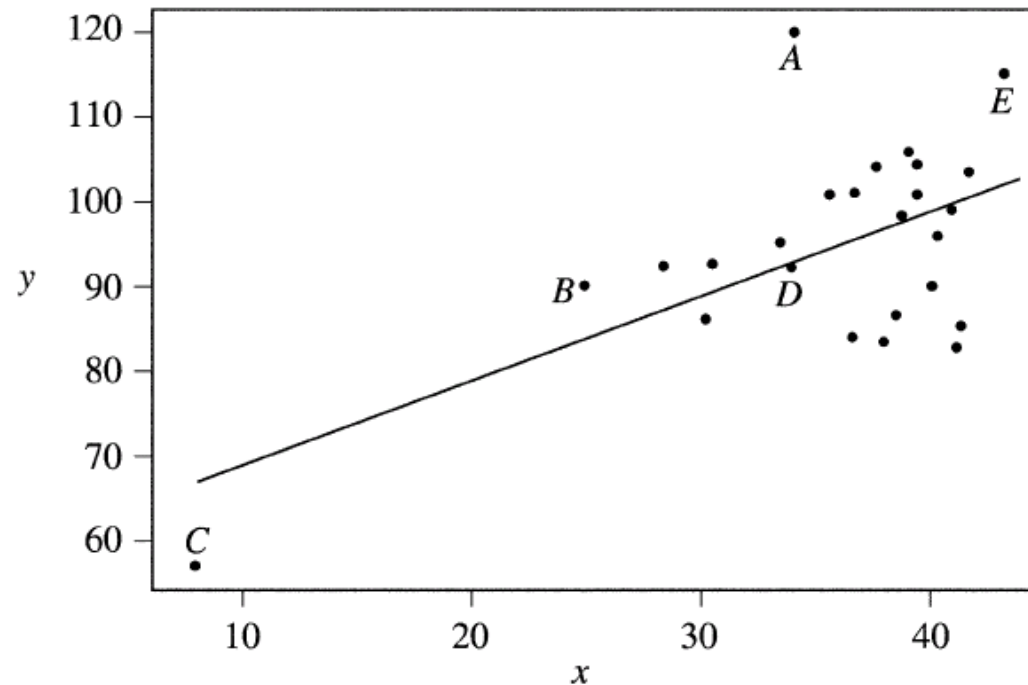
$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$b_1 = r \frac{s_y}{s_x}$$

$$s_{b_1} = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum (x_i - \bar{x})^2}}$$



1. In the scatterplot of y versus x shown above, the least squares regression line is superimposed on the plot. Which of the following points has the largest residual?

Problems with these 2 Formalisms

- Formulas are difficult for most students to interpret.
 - They are an algorithm fed into the student as computer.

Formulas and Tables

Students enrolled in the AP Statistics course should concentrate their time and effort on developing a thorough understanding of the fundamental concepts of statistics. They do not need to memorize formulas.

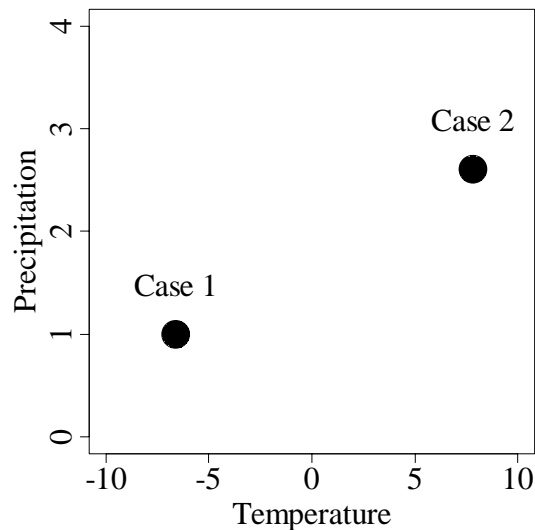
The following list of formulas and tables will be furnished to students taking the AP Statistics Exam. Teachers are encouraged to familiarize their students with the form and notation of these formulas by making them accessible at the appropriate times during the course.

- The scatter plot is clearly useful, but doesn't generalize well to multiple variables.

A Third Way: Geometry in Variable Space

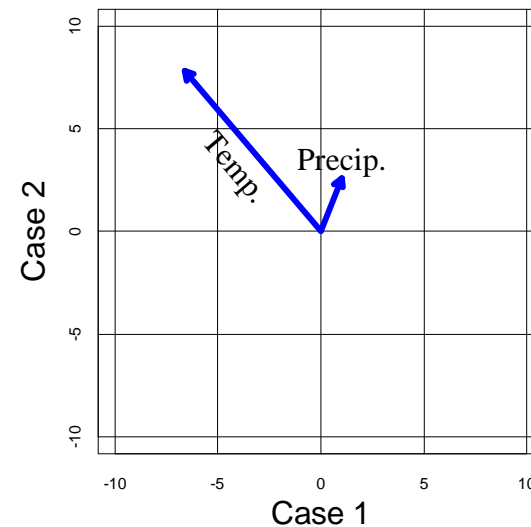
Case Space

- Conventional scatterplot
- Each **case** is one **point**.
- Each axis is one variable.
- We can show 2 variables, but many cases.



Variable Space

- Unfamiliar format
- Each **variable** is one **point**.
- Each axis is one case.
- We can show 2 cases, but many variables.

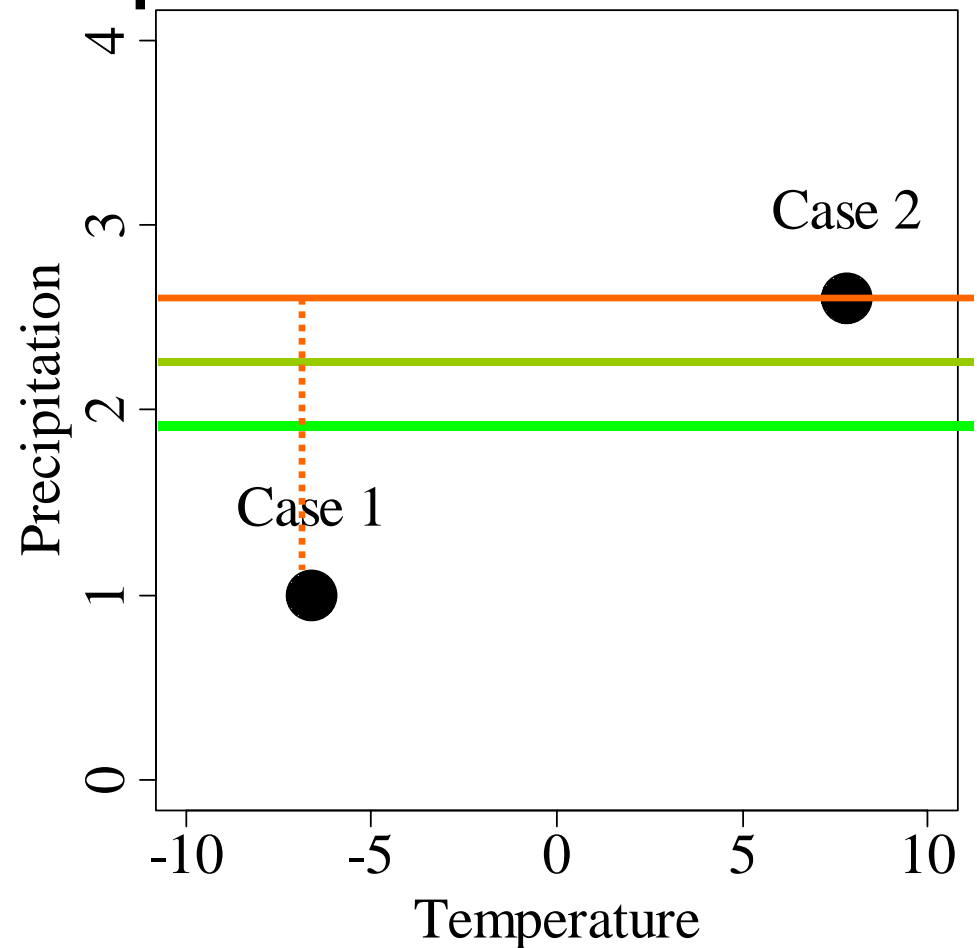


Pros and Cons of Variable Space

- To display N cases means using an N -dimensional space.
- Practical limit: displaying 2- or 3-dimensional spaces, but multivariate data sets with 2 or 3 cases are not so useful.
- Scatter plots can display large N in two variables: very useful. (And there are pairwise plots.)
- Some relationships are obvious in Variable Space.
 - We even use words that emphasize geometry: Colinearity and Orthogonality.
- Important statistical relationships can be visualized in two dimensions.
 - Practical calculations need to be performed in higher dimensions, but this is what computers are for. The student can understand what the computer is doing with only two-dimensional plots.

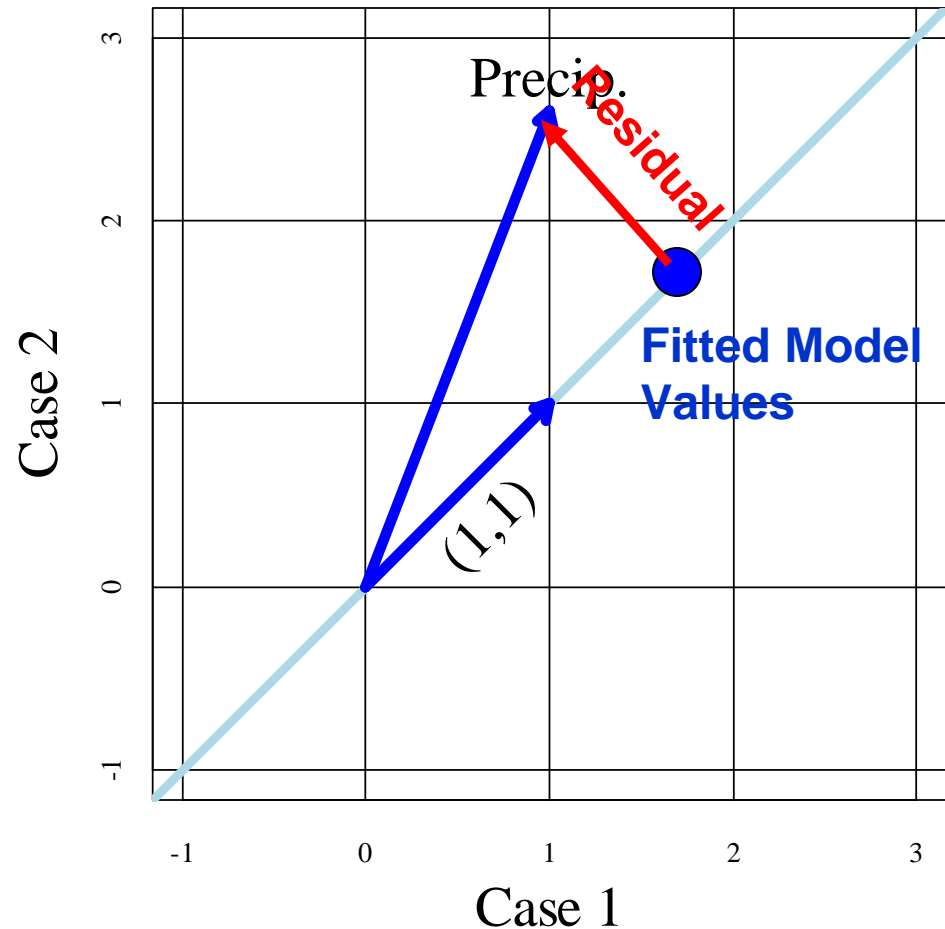
Fitting $Y \sim \text{constant}$ in Case Space

- In Case Space, find the function of the right form that comes as close as possible to the cases.
- Involves measuring multiple residuals, squaring and summing.



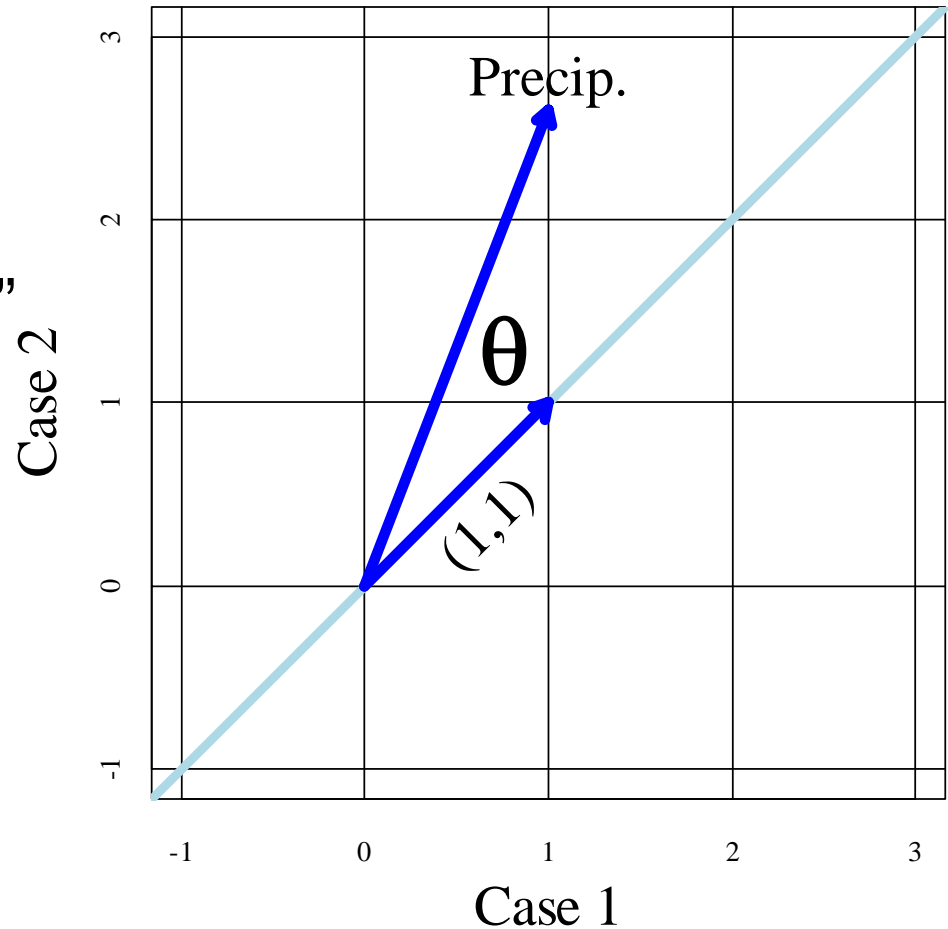
Fitting $Y \sim \text{const.}$ in Variable Space

- Model coefficient found by projection.
- Relationships that are clearly displayed
 - Residual is orthogonal to explanatory vector.
 - Partitioning of sums of squares: just the Pythagorean theorem.
- Moves students beyond the “mean” to general model coefficients.



Statistical Inference: A Geometrical One-sample t-test

- Null Hypothesis: Explanatory vector was picked at random.
- Test statistic: How “close” is the fitted point to Y . Closer is better.
- P-value: Simple geometry: $\theta / 90$ deg
 - no need for a t-table



The same idea applies for $N=3\dots$



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 49 degree co-latitude?

Correlation Coefficient

- Case Space
- Formula (AP course)

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

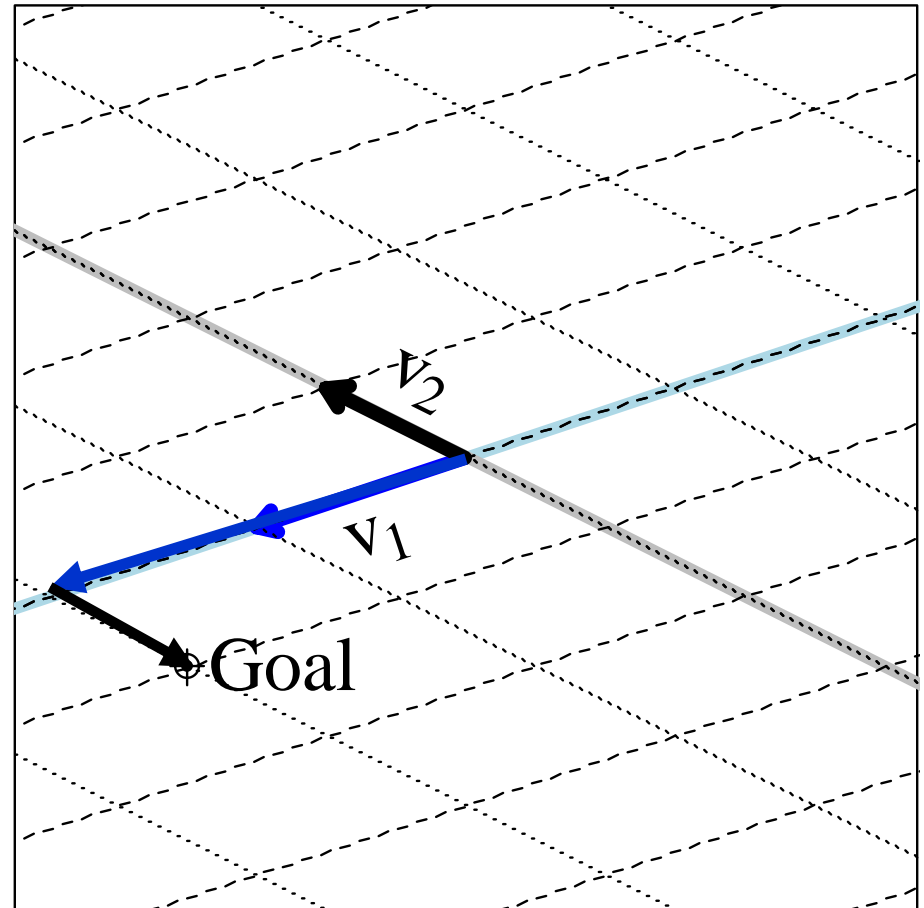
- Rotundity of scatter plots

Variable Space

- The angle between two vectors.
 - well ... $\cos(\theta)$
 - after mean is removed
- Easy generalization to multiple R^2
 - angle between fitted model values and Y

Multiple Regression

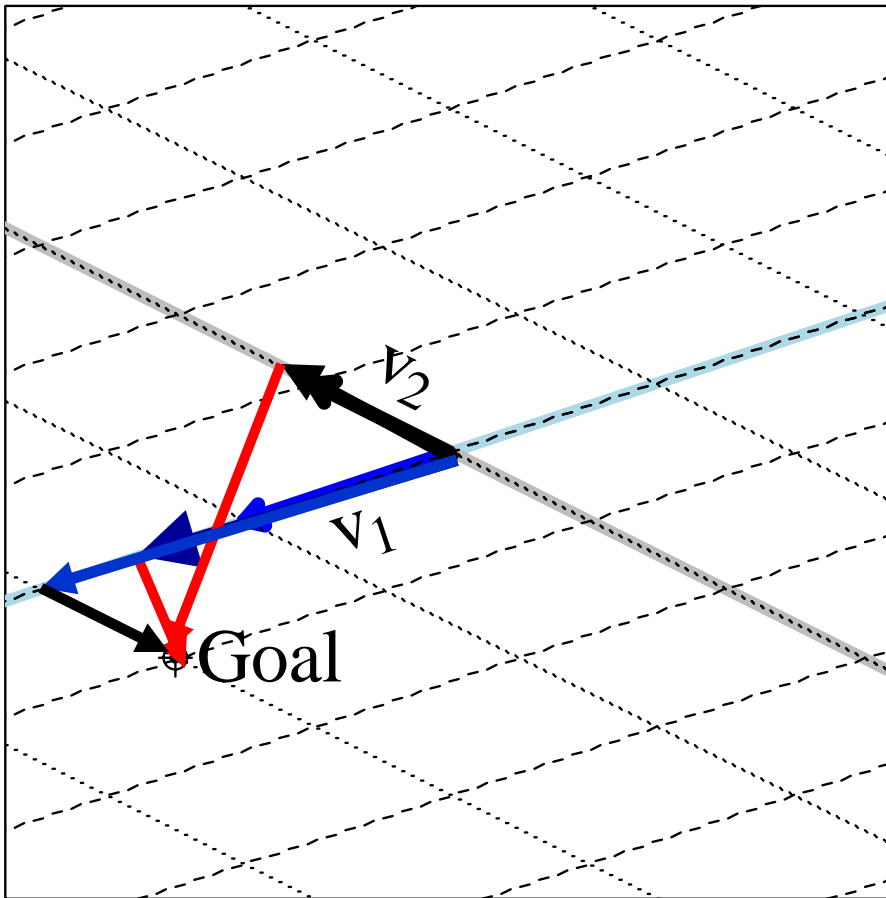
- Reach the goal point by walking in the directions of multiple vectors.
- With enough vectors, you can reach any goal.
- Vectors can be redundant
 - counting vectors leads to degrees of freedom



Colinearity

- When explanatory vectors are aligned, “funny” things can happen.
- Simpson’s Paradox has a simple geometrical interpretation.
- Why orthogonality is desirable.

Simpson's Paradox



V_1 only: positive coefficient

V_2 only: positive coefficient

V_1 and V_2 :

positive coefficient on V_1

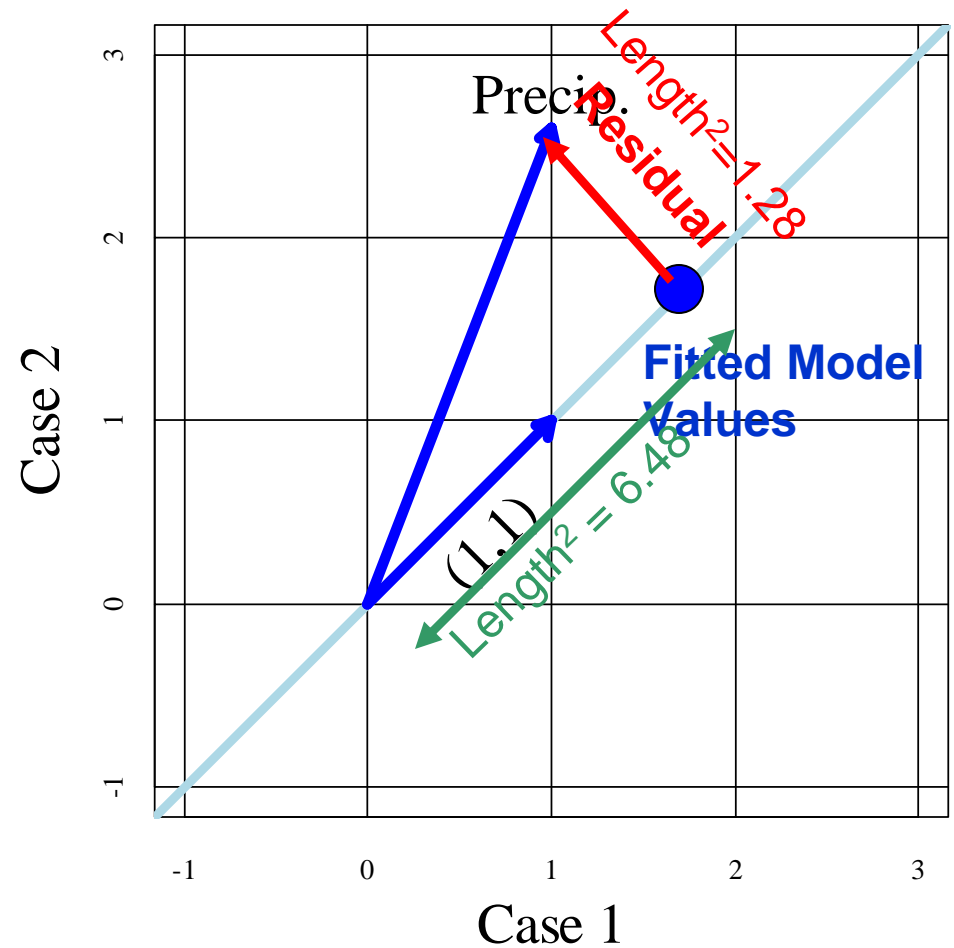
negative coefficient on V_2

Randomization and Orthogonality

- Random vectors (in high N) are almost always close to being orthogonal.
- Other ways to construct orthogonal vectors: e.g., blocking

ANOVA

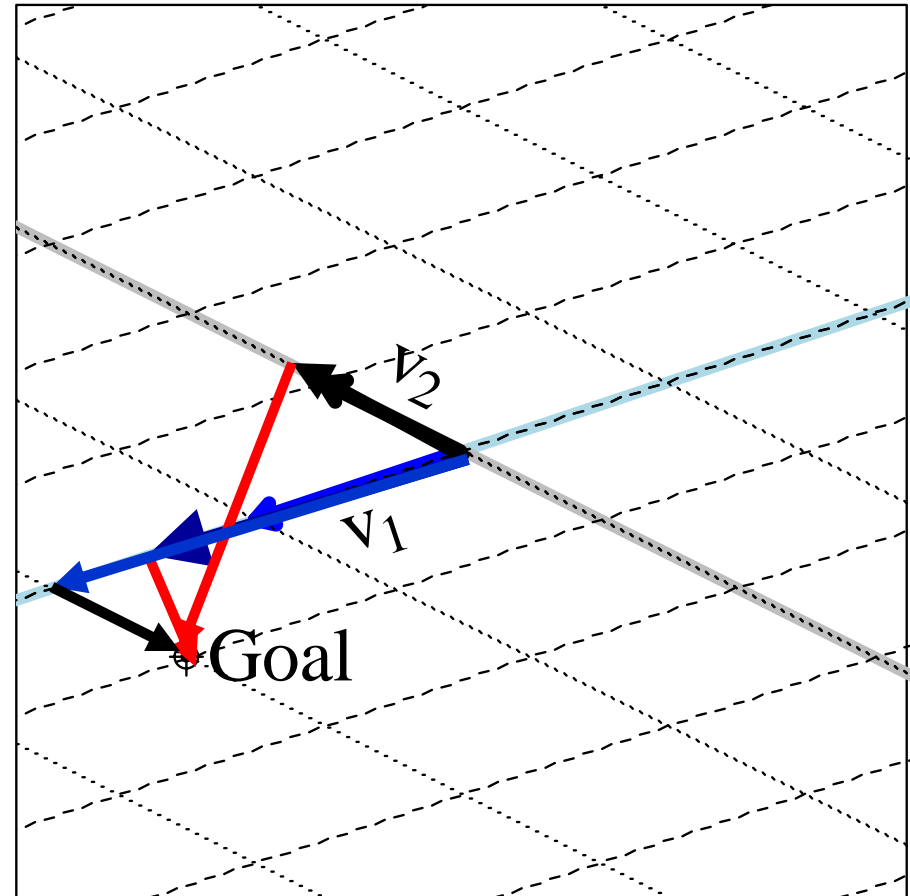
- Simple geometrical interpretation
- You can do the calculations with a ruler.



Term	D.F	SS	MS	F	p
(1,1)	1	6.48	6.48	5	0.27
Residual	1	1.28	1.28		

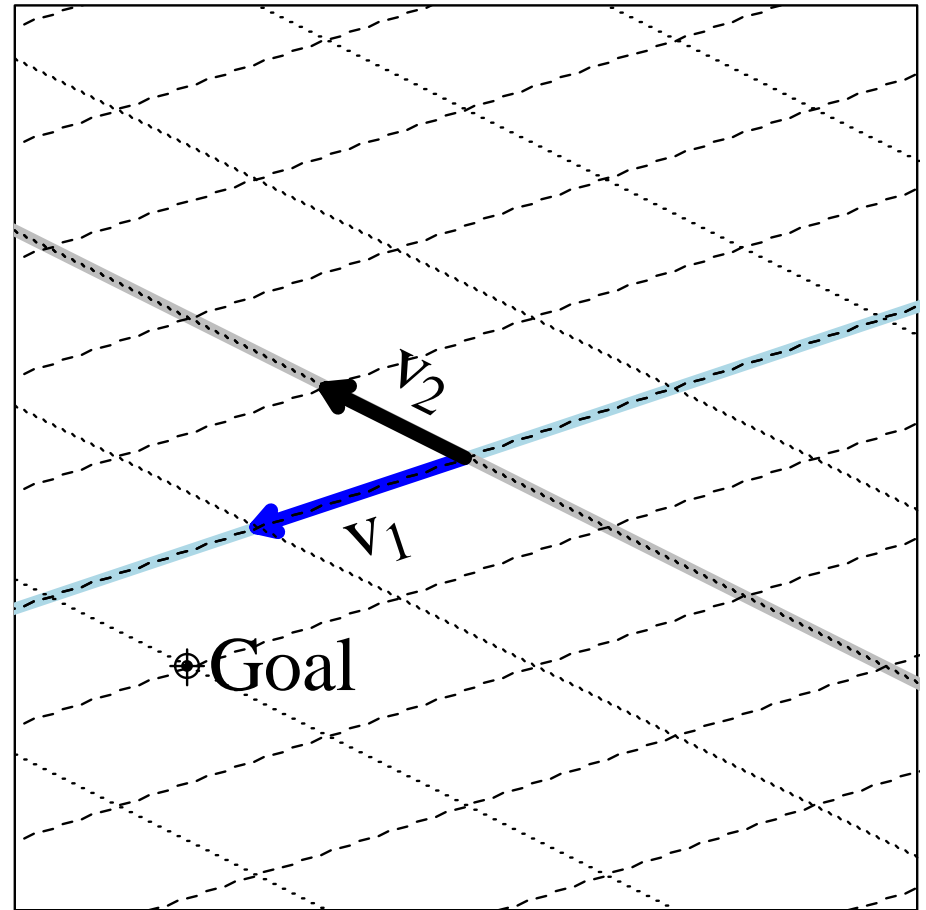
Multivariable ANOVA / ANCOVA

- Like regression, but add new variables in succession, rather than simultaneously.
- Order-dependence of ANOVA table (SS vs Adjusted SS) easy to visualize.
 - Why orthogonality eliminates this problem.
- ANOVA becomes a technique for looking at the role of a variable in the context set by others, not just a way to generate a p-value or answer “Are these groups different?”



3-Dimensional Space is Sufficient

- Although the data live in N dimensions, we can visualize the essential relationships in just 3 dimensional space.
- Model “plane” + orthogonal complement
 - v1: One set of explanatory factors (with it's associated DF)
 - v2: Another set of explanatory factors (and its DF).
 - The residual.
- Goal point hovers over the model plane. Residual points out of the paper.



Introduction to Statistical Modeling at Macalester

- presents statistics in a modeling framework
- uses geometry rather than algebra
- uses simulation and resampling rather than algebra
- exploits modern computation

Mainstream intro course. Approx. 100 students per year, mainly biology and economics majors, required for math majors. Pre-requisite: one semester of calculus.

They leave the course able to construct sophisticated models of complicated data sets.

Status of Project

- Large collection of exercises is available.
- Text using the geometrical approach being written (Draft available Summer 2007.)
- Looking for partners interested in adopting this approach at their own institution.
 - Many people tell me that “their students” can’t do this. If so, then the place to start is as a follow-up course to AP statistics, or a course for the mathematically inclined.