BioQuest Workshop, June 10-18, 2006

Exploring Complex Data Sets

Daniel Kaplan, Macalester College

Notes on Teaching Multivariate Statistical Modeling to Introductory Students

The title of this workshop refers to "complex data sets." Of the several dictionary definitions of "complex," the one that seems most directed to our meaning is

> A group of obviously related units of which the degree and nature of the relationship is imperfectly known.[1]

It's natural to think about the "units" of the dictionary definition as the cases in a data set: our experimental units, our samples, the rows of a spreadsheet table. But for thinking about statistics, I believe this is wrong. A dataset is complex not because it has lots of rows, but because it has lots of variables. It is the relationships among the variables that we seek to elucidate through statistical analysis.

These notes outline a new approach to teaching introductory statistics that I have been developing at Macalester College over the past five years. The notes are intended for readers who have already studied statistics in the conventional way. I assume you already know about hypothesis tests, t-tests, simple linear regression and such and that you have heard about analysis of variance. In the course I teach, we make no such assumptions. But these notes are not the course notes. They are notes *about* the course and provide a concise description of the approach.

My goal in developing the new approach is to give natural science and social science students the ability to use statistics in authentic and legitimate ways to carry out work in their chosen fields. Often, that work will involve complex data sets; sets with multiple variables. This goal might seem uncontroversial, but it contrasts with the stated goals of a typical introductory statistics course of giving students a general appreciation for statistical reasoning.

A typical course covers many worthwhile topics: sampling, randomization, simple descriptive statistics, inferential tests such as the t-test and chi-squared test, simple linear regression, p-values. When taught well, such an introductory

[1] Mirriam-Webster online dictionary

course can bring students to understand how to reason about uncertainty in quantitative ways and gives them tools for studying simple relationships between variables.

The problem is the word "simple." The questions that the introductory tools allow students to answer are simple: Are these two groups different? Are these two variables correlated? These questions are insufficient for much scientific work, which commonly involves multiple factors (not just two) that are related in complicated ways. If our goal really is to allow students to work with complex data sets, we should think carefully about how we give them the skills needed to do so. Typically, a student who plans to go on in scientific research will have to take a "methods" course, sometimes in graduate school, that teaches the disciplinary techniques, often multivariate, used in their field. This has obvious consequences for the type of research work that students can carry out as undergraduates. It also means that the introductory course is not strongly relevant to the student's chosen field.

The new approach is based on several principles:

- From the very beginning, students should be able to use statistical tools to study complicated relationships among multiple variables. Students should not have to distort the questions they want to ask to fit into the are-these-two-groups-different framework. Statistical techniques should provide insight to genuine questions, not just a formal validation of the answers to artificially simple questions.

- Strong connections should be drawn between statistics and other areas in science. A common theme throughout science is modeling and approximation. Statistics has the potential to enhance strongly a student's understanding of the process of modeling. Or, to take a small example, the $\sqrt{N}$ that appears so often in statistics is closely related to the physical phenomenon of diffusion.

- The computer is a basic tool of scientific work students should be expected to learn to use it well and to rely on it. One consequence of this is that we can teach in ways that exploit the availability of computation rather than overlapping with it. We no longer need to focus on algorithms that can be carried out by hand or even to think of the computer as a labor-saving device. The computer allows us to do new things that would be unimaginable to do by hand. It frees us to think about statistical processes at a higher level.

- Related to the above, it is no long necessary to present statistics in terms of formulas. In the past, statistical formulas have played two roles: they have presented an algorithm for a calculation (e.g., "calculate the t-value using the formula") and they have been used to provide a notation for describing the statistical concepts. As for the first purpose, computers have now freed us from the need always to describe algorithms in human readable form. As for the second, it has never been very effective for most

students. Aside from a few specialists (such as mathematics and statistics professors!), most people have a hard time reading or understanding formulas. Although the formulas provide one way of presenting concepts, there are other ways that can be much more effective for some people and that can provide new insights.

- If we simply replace formulas that students understand with computer routines that students don't understand, we lose the opportunity to let students think creatively about statistics and to be able to check the results they get from the computer for reasonableness. We need to give students a formalism that let's them think about statistics. In the Macalester course, we make heavy use of a geometrical approach to understanding statistical modeling. This let's us avoid the linear-algebraic difficulties of multivariate modeling (transposes! inverses! singularity! pseudo-inverses!)

  We use a simulation/resampling approach to let us consider sampling distributions in a way that is intuitively accessible to students.

- Despite the importance of learning to use the computer, students need to be able to reason about statistics without depending on the computer. For example, it sometimes happens that including a new variable in a model will completely change the influence of a previous variable. This is easily seen by examining the coefficients of a computer print-out, but it's important for students to understand in what circumstances this sort of thing can happen and what they can do to avoid the ambiguity of interpretation that this introduces. This connects directly to experimental design.

- It is more important to provide students with a simple, general framework for interpreting data than to lead them through a zoo of specialized tests. Specialized tests are for specialists. In the introductory course we seek to create generalists who can understand how specialized tests fit in and when they are worthwhile. Some examples of specialized tests are those involving very small N, the unequal variance t-test and the variety of non-parametric tests such as the rank-sum test.

ISM at Macalester. At Macalester, we teach Introduction to Statistical Modeling as a first-level course. Our main clients are the biology and economics departments, both of which require ISM for their majors. (Psychology majors can also take ISM.) All ISM students are required to have had a semester of calculus; AP calculus will do. We prefer if they have taken our Applied Calculus course, which is unusual in being about multivariate modeling and by design ties directly to the Statistical Modeling course.

For a school or for departments that do not require any calculus, the statistics course would need to be somewhat modified. The main change would be the need to cover the material in Section 2 on linear approximating functions. (The geometrical and resampling approaches that are at the core of the course, do

not call on calculus.) One possible approach would be to treat Introduction to Statistical Modeling as a *second* statistics course, so that the time that we spend in ISM teaching the framework of hypothesis testing or probability could be devoted instead to linear approximating functions.

This second-course approach might also be effective in dealing with a source of institutional resistance to ISM: the claim that students at College X would not be able to learn the "advanced" material taught in ISM or that they will suffer an irretrievable loss by our omission of some of the standard topics in introductory statistics (e.g., the unequal variance t-test). We certainly encountered this claim at Macalester. Our experience at Macalester have been very positive, but we are aware that circumstances differ from school to school, both in terms of the background of students and the level of support provided by departments. Our initial experiment with ISM got off the ground because of the strong participation, encouragement, and commitment of our biology department. Importantly, our math department was willing to risk something new and devote resources to course and faculty development.

Statistical software. All students who take ISM at Macalester are expected to become proficient with statistical software. We use the R statistics package: a free, professional-level package that runs under Windows, Mac, and Linux. There is undeniably a learning curve to R; it will take several hours of experience (a couple of hours of which is in class) for students to become proficient. Our experience is that this investment of time is well worthwhile. We also use R in the Applied Calculus class, so our ISM sections are well seeded with students who are already comfortable with the package.

It would be possible to teach most of ISM with another package, such as SPSS or STATA. It isn't clear that the simulation and resampling components could be easily implemented in such packages. The textbook for ISM is being written in a manner that is independent of any computer package; just the exercise sets make use of software. So, if there is interest, it should be possible to port the course to another software package.

1 Some Example Data & Projects

To illustrate the approach, in these notes we'll focus on three examples. One dataset is very small and simple and is used to illustrate some modeling techniques. The other two involve more cases and more variables. We will use them for projects.

Some readers may have data sets of their own that they would like to work on using the modeling techniques presented here, or that they would like to include in a course (or a statistics textbook!). If so, do come talk to me!

To use the data, you need to start the R software and loading in a single file, `ISM.RData`, that is pre-formatted with the data and some custom software. You will find the `ISM.RData` file at `www.macalester.edu/~kaplan/ISM.RData`;

copy it to your own computer. The process of loading the file is shown in Figure 1.

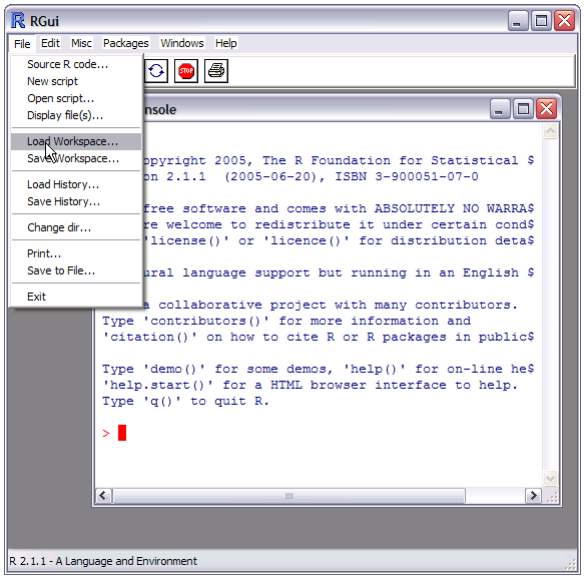


Figure 1: Starting the R statistics package and loading in a workspace with predefined data.

1.1 Swimming records

The data table in the variable `swim`, holding the data from the spreadsheet file `swim100m.csv`, contains world-record times for swimming the 100 meter freestyle. The variables are the time (in seconds), the year in which the record was set, and the sex of the swimmer.

1.2 Height as a heritable trait

In the 1880's, inspired by the recent work of Darwin, Francis Galton was developing ways to quantify the heritability of traits. As part of this work, he collected data on the heights of adult children and their parents.

The data were transcribed by J.A. Hanley, who has published them at `http://www.medicine.mcgill.ca/epidemiology/hanley/galton/`. [2]

The variable `galton` (equivalent to the spreadsheet `galton-heights.csv`) contains most of Galton's recorded measurements.

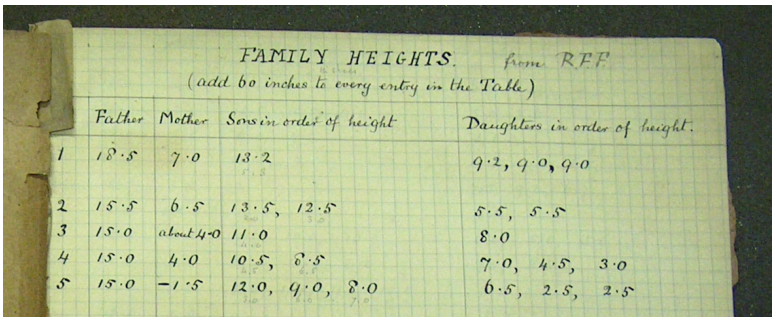
[2] See J. A. Hanley, "Transmuting" women into men: Galton's family data on human stature, is published in The American Statistician, 1 August 2004, vol. 58, no. 3, pp. 237-243. The photograph of Galton's notebook is also from Hanley.

family	father	mother	sex	height	nkids
1	78.5	67.0	M	73.2	4
1	78.5	67.0	F	69.2	4
1	78.5	67.0	F	69.0	4
1	78.5	67.0	F	69.0	4
2	75.5	66.5	M	73.5	4
2	75.5	66.5	M	72.5	4
2	75.5	66.5	F	65.5	4
2	75.5	66.5	F	65.5	4
3	75.0	64.0	M	71.0	2
3	75.0	64.0	F	68.0	2

and so on

In the reformatted table, each child is one case. The variables are the child's height, sex, the number of children in the family, and the heights of the child's father and mother. Entries were deleted for those children whose heights were not recorded numerically by Galton, who sometimes used entries such as "tall", "short", "idiotic", "deformed" and so on. There is a unique code for each family, which is stored as a categorical variable.

Galton's original format was different and possibly would make more sense intuitively to a student. Contrasting the table and the original format helps to drive home the modern notation that a case is one row in a table, a variable is one column in a table.



FAMILY HEIGHTS. from REF

(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5

Project. Following Galton, we want to describe the extent to which height of a child can be ascribed to height of the parents. Of course, there are other variables that may play a role: the child's sex, the family's nutrition and other environmental factors, the number of children in the family (which may reflect the resources available to each child, or might reflect the health of the parents.)

Galton developed the correlation coefficient to help with his study of data such as these. He did not have today's multivariable techniques. To take into account both the father's and mother's height, he computed a "mid-parent," a weighted average of the father's height and 1.08 times the mother's height. We can include both father's and mother's height as separate variables.

The data are available in the R variable `galton` that is contained in the `ISM.Rdata` file, or, equivalently, in the spreadsheet `galton-heights.csv`. Once you have loaded the `ISM.Rdata` file, you will have the data set available to use. To remind yourself of the names of the variables, use the command

```
> names( galton )
```

The basic computational tool we will use is linear modeling. Models can be constructed with statements like this:

```
> mod = lm( height ~ sex + father + family, data=galton)
```

Questions.

1. How much variability is there from child to child? One way to characterize this is by the mean-square of the residuals of the simple model `height ~ 1` A more familiar term for this mean-square is the variance — the square of the standard deviation.

2. To what extent does the child's sex account for height? That is, how does including the `sex` variable in the model reduce the mean-square of the residuals.

3. How much does the father's height account for the child's height? How about the mother's height? Does including both of the parent's heights improve things further? Is there any indication of an interaction between sex and parent's height? Of an interaction between father and mother's heights? (Such an interaction term would say that the influence of a father's height will be different for different mother's height.)

4. Define a new variable containing Galton's mid-parent. While we're at it, we can also try the straight average of the parents' heights.

   ```
   > midparent = (galton$father + 1.08*galton$mother)/2
   > aveparent = (galton$father + galton$mother)/2
   ```

   Does the `midparent` or the `aveparent` capture as much of the variability in childrens height as father and mother a two variables?

5. Does the number of children in a family have an influence on the child's height?

6. As a proxy for all the other environmental factors that influence height, we can use the `family` variable. Do note that there is a heavy redundancy between `family` and the variables `father`, `mother`, and `nkids`. The reason is that in each family in this data set, there is only one father, one mother, and a set number of kids. Since the `family` variable could be used to make an exact model of any of these variables (try it!), there is redundancy. We can say that father and mother are nested in family.

   In order to see the influence of father, mother, or number of children, those terms will have to come *before* family in the model specification.

7. For an example of a pitfall of statistical modeling, construct the following two models:

```
> mod1 = lm( heights ~ sex + father + mother, data=galton)
> mod2 = lm( heights ~ sex + father + mother + father:mother, data=galton)
```

   Use the `summary` command to look at the coefficients of each model and the p-value that describes whether we can reject the null that the coefficient is zero. Adding in the interaction term causes all significance of the parent's influence to disappear. Yet the ANOVA report shows something different. Why do you think this is?

1.3 Birthweight and smoking

The variable `birth`, containing the data from a spreadsheet file `birthweight.csv`, contains data from the Child Health and Development Studies used to explore the link between maternal smoking and infant health. As described in Nolan and Speed,[3] the entire dataset, of which this file is a sample, includes all pregnancies that occurred between 1960 and 1967 among women in the Kaiser Foundation Health Plan in Oakland, California. The subset is restricted to male babies who survived at least 28 days after birth, and contains information about the baby and its mother and father, including variables such as length of gestation, baby weight, the weight, height, education, income, and race of the parents and the extent to which the mother smoked.

Project: Your job is to model the babies' weight at birth. Nolan and Speed were interested particularly in how birthweight depends on the mother's smoking; is there a direct dependence or is is mediated by the length of gestation?

In the spreadsheet file, the data have stored as numerical codes even if they are nominal data. This was a common style a decade or two ago, and you are likely to encounter data in this format. Even missing data was encoded with a simple number, say 99. This can cause serious problems when that number is also a legitimate data value. For this reason, the modern style of encoding missing data as NA is much to be preferred.

In the `birthweight.csv` file, we have translated all missing data to NA, but other than that we have left the coding in its original, numerical format. The significance of this is that, for those variables that are nominal, you should use the `as.factor` command so that the data are not inappropriately taken to be numerical. For instance:

```
> b = read.csv('birthweight.csv')
> lm( wt ~ as.factor(smoke), data=b)
```

[3] D. Nolan and T.P. Speed, *Stat Labs: Mathematical Statistics Through Applications*, Springer, 2001

Of course, variables such as the mothers' ages, weights, and so on, should properly be taken as quantitative data. You will notice when you have inappropriately used a nominal variable as quantitative when you are surprised to get a single coefficient for that variable, rather than one coefficient for each level.

Why didn't we bother to recode the data so that nominal data took on the levels described in the codebook? We want you to learn to pay attention to the difference between a quantitative variable and a nominal variable that has been coded with a number.

Variables in the data file:

date birth date where 1096=January 1,1961

gestation length of gestation in days

sex infant's sex 1=male 2=female 9=unknown. (All the cases in this data set are males.)

wt birth weight in ounces (999 unknown)

parity total number of previous pregnancies including fetal deaths and still births, 99=unknown

race mother's race 0-5=white 6=mex 7=black 8=asian 9=mixed 99=unknown

age mother's age in years at termination of pregnancy, 99=unknown

ed - mother's education

```
0= less than 8th grade,
1 = 8th -12th grade - did not graduate,
2= HS graduate--no other schooling ,
3= HS+trade,
4=HS+some college
5= College graduate,
6&7 Trade school HS unclear,
9=unknown
```

ht mother's height in inches to the last completed inch 99=unknown (I think this means "rounded down to the nearest inch")

wt.1 mother prepregnancy weight in pounds, 999=unknown

drace father's race, coding same as mother's race.

dage father's age, coding same as mother's age.

ded father's education, coding same as mother's education.

dht father's height, coding same as for mother's height

dwt father's weight coding same as for mother's weight

marital Marital status.

```
1=married,
2= legally separated,
3= divorced,
4= widowed,
5= never married
```

inc family yearly income in $2500 increments 0 = under 2500, 1=2500-4999, ..., 8= 12,500-14,999, 9=15000+, 98=unknown, 99=not asked

smoke does mother smoke? 0=never, 1= smokes now, 2=until current pregnancy, 3=once did, not now, 9=unknown

smoker does the mother smoke now?

time If mother quit, how long ago? 0=never smoked, 1=still smokes, 2=during current preg, 3=within 1 yr, 4= 1 to 2 years ago, 5= 2 to 3 yr ago, 6= 3 to 4 yrs ago, 7=5 to 9yrs ago, 8=10+yrs ago, 9=quit and don't know, 98=unknown, 99=not asked

number number of cigs smoked per day for past and current smokers 0=never, 1=1-4 2=5-9, 3=10-14, 4=15-19, 5=20-29, 6=30-39, 7=40-60, 8=60+, 9=smoke but don't know, 98=unknown, 99=not asked

id Identification number (not useful for modeling)

pluralty 5= single fetus (all cases are the same)

outcome 1= live birth that survived at least 28 days. (All the cases are this way)

2 Linear Approximating Functions

Most of a conventional introductory statistics course is built around some very basic descriptions of a single variable: the mean, the count, the proportion. From these familiar ideas, some less-familiar descriptions are constructed to be used in statistical inference: the standard deviation and standard error of the mean, χ^2, and so on.

In ISM we introduce, at a very early stage, an additional way of describing data that reflects the relationship between variables. The fundamental distinction is that one variable is selected as the *response* and others as the explanatory variables. We will model the response as a function of the explanatory variables. For convenience in notation, we'll call the response variable z and the explanatory variables as x and y. Of course, in general there will be more than two explanatory variables, but we can illustrate all the concepts we need in ISM with just two explanatory variables; students have no trouble generalizing to more. The general modeling scheme is $z = f(x, y)$.

Almost all students at the college level are familiar with the linear function $z = mx + b$. They know about slopes and intercepts, they know what the graph of the function looks like. They often don't know that this is a general purpose approximation to a wide range of functional relationships.

Where things become new for many students (at least, those who haven't taken our Applied Calculus course) are functions of two variables. An algebraic form that is linear in both x and y is $z = a + bx + cy$. We work students through this simple form, showing the graph as an inclined plane, showing the contour-plot form of this graph, plugging in values for x and y, showing that the slope of the function — rise over run — depends on which direction in x, y-space one considers. The basic lesson is that for this form of function, the value of z depends on x in the standard slope-times-run format they are familiar with, and depends on y in a similar way but potentially with a different slope.

Important special cases of this function are these:

$b = 0$ where z doesn't depend on x

$c = 0$ where z doesn't depend on y

$b = 0$ **and** $c = 0$ where z doesn't depend on either

The point is that looking at the coefficients can tell us which variables are related and that a coefficient value of zero indicates no relationship. Later on in the course, this will be important in interpreting hypothesis tests.

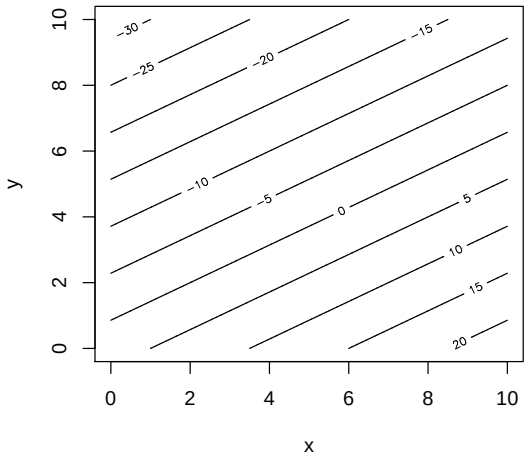


Figure 2: Contour plot of a function $f(x, y) = a + bx + cy$. The spacing between contours is constant, independent of x and y. That is, the slope with respect to x is independent of y, and vice versa.

Few students have difficulty with this, particularly when presented with simple examples such as z being income, x being hours worked at one job, y being hours working at another job. b and c are the wage rates at the different jobs, while a is income from gifts, etc.

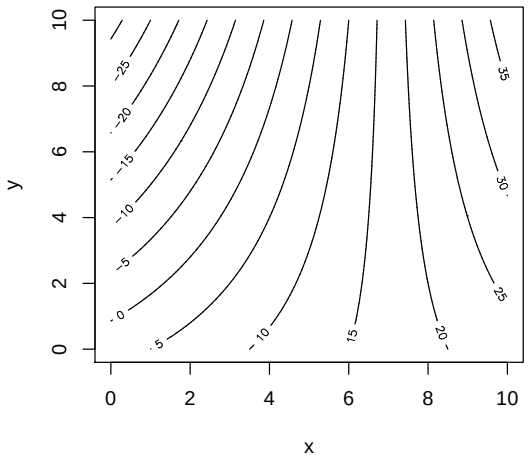


Figure 3: Contour plot of a function with an interaction term $f(x, y) = a + bx + cy + dxy$. The spacing between contours depends on x and y. That is, the slope with respect to x depends on y, and vice versa.

The next important model form is similar, but is linear in *each* of x and y. This is $z = a+bx+cy+dx \cdot y$. What's new here is the term $dx \cdot y$: the coefficient d is multiplying a simple product of the two variables. There are many different names for this term: in chemistry (where z might be the rate of formation of a species) it reflects the Law of Mass Action; in mathematics it is called a "bilinear term," in statistics it is called an "interaction term." The statistical name brings to mind various classical models that depend critically on such a term, for instance the famous Lotka-Volterra predator-prey model where it reflects the rate at which predators encounter prey, or the SIR epidemic models where it reflects the rate at which infected people meet susceptible people.

A concrete example of an interaction term is given by a data set on world-record times in the 100m freestyle swimming race. The world record has gotten faster over the years. Depending on what terms one includes, one can model different records for men and women or, with an interaction term, different slopes for men and women, because the dependence on year depends on sex.

We work with the interaction term in very concrete ways. It allows us simultaneously to model a linear dependence on x where the slope depends on the value of y. We work through examples numerically to show how, when there

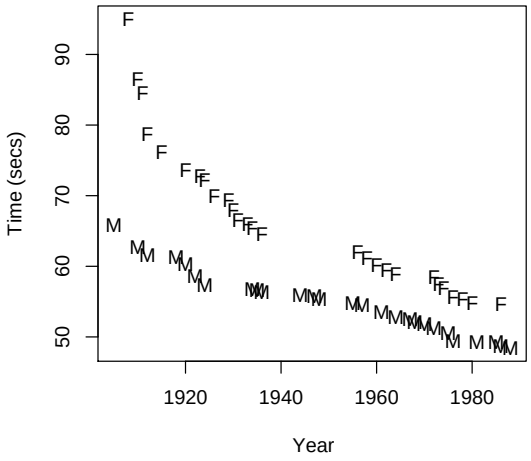


Figure 4: World-record times for swimming 100m over the years for males and females.

is an interaction term, the dependence on x depends on y. A useful exercise is to construct word models of a situation and translate them into a choice of terms. For example, suppose we want to model how fast a bicycle goes depending on the slope of the terrain x and the gear y. In any given gear, that is, for fixed y, speed depends on slope x: positive slope makes the bike go slower, negative slope makes it go faster. So there needs to be a term bx (with $b < 0$). On a flat road, the bike goes at a speed $a + cy$ — it depends on the gear — so we need the terms a and cy. Finally, the way the speed depends on slope itself depends on the gear: we would go very slowly uphill if we were in a low gear. This means there is an interaction between slope and gear: the dxy term.

A sign of no interaction between x and y in determining z is when the coefficient d on the dxy term is zero.

If one wants to go further — but there is no requirement to do so — one can generalize the function $f(x, y)$ to higher-order polynomials. We do this in Applied Calculus, working things out to quadratic terms because we want to study optimization. When talking about functions such as $z = a + bx + cx^2$, we emphasize that the x^2 term is an interaction of x with itself: the dependence of z on x depends on x.

Interpreting the coefficients of models requires that students understand the nature of units, and that they know what a partial derivative is. Neither of them is difficult, but college mathematics curricula tend to avoid units entirely and introduce partial derivatives in an overly symbolic manner that isn't even reached until the third semester of college calculus. Our Applied Calculus course

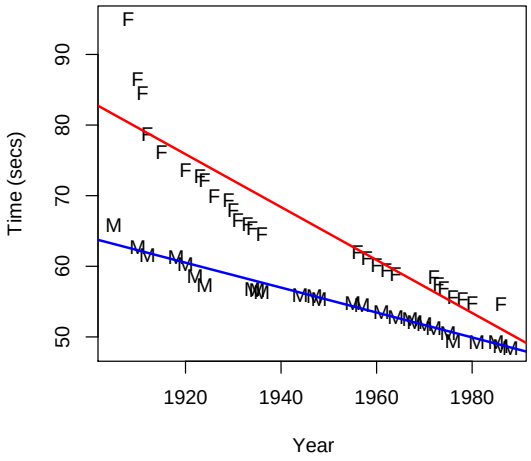


Figure 5: A model of world-record swim times that includes an interaction between year and sex. The interaction term represents the dependence of swim time on year as it vary with sex, or, equilvalent, represents the difference between the sexes as it changes over the years.

takes units seriously (and scientists need to understand units). We emphasize the meaning of a partial derivative: how a response variable changes as one of the explanatory variables is changed but *holding all the other variables constant.* There doesn't need to be any algebra involved; it doesn't even need to be a "calculus" topic.

By focusing on R^2 rather than coefficients, statistics courses can sidestep the need to talk about units and partial derivatives. But it's important to be able to talk about the *strength* of relationships in quantitative terms. To give an example, recently there has been a small controversy about the use of sunscreen. By reducing exposure to sunlight, sunscreen can help reduce the risk of skin cancer. But it also reduces the production of vitamin D which can, in turn, increase the risk of other types of cancer. Knowing the R^2 of these relationships won't help in figuring out if it's worth going out without lotion in the winter; you need to know *how much* the risk is changed in each case by going without lotion.

Exercise 1

In Lotka-Volterra type models of population growth under predation, one writes the rate of change in the population of prey as a function of the populations of both predator and prey. Imagine that we have rabbits R and foxes F, and we

create a model for the rate of change of rabbits in time as

$$\frac{dR}{dt} = f(R, F) \approx \alpha R - \beta R \cdot F.$$

Which of these is an interaction term? Suppose that for some reason rabbits become better at evading their predators. Which coefficient would this change?

Exercise 2

Consider the birthweight data set. To study the relationship between smoking and weight at birth we will examine a bi-linear model of the form

$$W = a + bG + cS + dG \cdot S$$

where W is the birthweight in ounces, G is the length of gestation in days, and S is a variable that indicates whether the mother smoked or not during pregnancy.

1. Describe, in everyday language, what each of the coefficients, a, b, c, and d stands for.
2. Suppose that smoking simply reduces birthweight by, on average, 25 grams. How would this show up in the coefficients?
3. Suppose that smoking causes babies to grow slower per week of gestation. How would this show up in the coefficients?
4. Suppose that the only effect is that smoking shortens gestation, but that babies of smokers grow in the same way per week of gestation as babies of non-smoking mothers. How would this show up in the coefficients?

Exercise 3

To describe the relationship among the variables in the Galton height data, we are interested in a model of the form:

$$H = a + bS + cF + dM + eF \cdot M + fS \cdot M$$

1. Which coefficient characterizes the influence of mothers on their children's adult height?
2. Suppose that tall fathers potentiate the influence of mothers. That is, a mother's height has more influence if the father is tall. How would such an effect show up in the coefficients?
3. Suppose that mothers have the main influence on their sons' heights, but that both fathers and mothers have an equal influence on daughters' heights. (For instance, suppose height were an X-linked trait.) How would this show up in the model coefficients?

3 Model Fitting

When students understand the form of models, they can start to fit models to actual data.

We take seriously the organization of data, requiring students at all times to work with data in a tabular form, where rows are *cases* and columns are variables. Every data set that we work with comes in the form of a spreadsheet. There are valuable lessons to be learned about what constitutes a case. In ISM at Macalester, we even talk briefly about relational database operations, but this is not essential to the course.

Software makes it straightforward to find the best-fitting coefficients on models so long as one knows how to select the model form and specify this form to the software.

In R, the notation for specifying a model is straightforward. If we are interested in the model $z = a + bx + cy$, where we want to find the best fitting coefficients a, b, and c, we write `z ~ 1 + x + y`. To include an interaction term dxy, we can write `z ~ 1 + x + y + x:y`

Here is an example, where we want to model swim record times by year and sex. To follow the example, make sure to start up R and load the data as described on page 5.

```
> names(swim)
[1] "year" "time" "sex"
> mod = lm( time ~ 1 + year + sex, data=swim)
> mod
Coefficients:
(Intercept)         year         sexM
   601.5320      -0.2751     -10.0530
```

The coefficient on year says that the record time has been decreasing by 0.275 seconds per year. The coefficient labeled `sexM` says that record times for men are about 10 seconds less than for women. The intercept coefficient, 602 is the swim time for women in a hypothetical "year zero," that is, when the year variable is zero. In that year, men would have had a swim time of $602 - 10$.

A simple summary of the quality of the model is provided by the coefficient of determination, R^2. This coefficient is the ratio of the variance of the fitted values of the response — that is, the ideal values that come from the model formula — to the variance of the actual response data.

```
> var(mod$fitted)
[1] 85.0342
> var(swim$time)
[1] 100.9094
> var(mod$fitted)/var(swim$time)
[1] 0.8426787
```

A more comprehensive report is also readily available.

```
> summary(mod)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 601.5320    42.5768  14.128  < 2e-16
year         -0.2751     0.0219 -12.560  < 2e-16
sexM        -10.0530     1.1150  -9.016 3.89e-12

Residual standard error: 4.062 on 51 degrees of freedom
Multiple R-Squared: 0.8427,     Adjusted R-squared: 0.8365
F-statistic: 136.6 on 2 and 51 DF,  p-value: < 2.2e-16
```

The standard error can be used to construct a confidence interval on each individual coefficient. The p-value relates to the null hypothesis that the corresponding coefficient is zero. Although we won't do so in these notes, in class we spend considerable time talking about confidence intervals and hypothesis tests. We use simulation and resampling to show where the standard error and p-value come from, as we'll illustrate in Section 5.

Rather than worrying about how to compute a standard error, we study the properties of the standard error, for example how it depends on the number of cases N. Again, we'll illustrate this in Section 5.

Another important description of a model is analysis of variance (ANOVA):

```
> summary(aov(mod))
            Df Sum Sq Mean Sq F value    Pr(>F)
year         1 3165.6  3165.6 191.880 < 2.2e-16
sex          1 1341.2  1341.2  81.297 3.893e-12
Residuals   51  841.4    16.5
```

Before the students see this report, they have been taught about sums of squares and such. In these notes, we'll outline the approach in Section 4. There's no good reason, in our approach, to distinguish between one-way and two-way ANOVA, or even ANCOVA.

In the swim-record data, it's clear that the dependence of time on year varies between men and women. Thus, an interaction term `year:sex` is called for. ANOVA shows that the term is statistically significant.

```
> mod2 = lm( time ~ year + sex + year:sex, data=swim)
> summary(aov(mod2))
            Df Sum Sq Mean Sq F value    Pr(>F)
year         1 3165.6  3165.6 314.454 < 2.2e-16
sex          1 1341.2  1341.2 133.230 1.037e-15
year:sex     1  338.0   338.0  33.579 4.555e-07
Residuals   50  503.3    10.1
```

This is clearly not the *best* possible model. The linear form, even with an interaction with sex, doesn't capture the pattern of the data very well. The situation becomes a little difficult since the nature of world-record data is that time will never increase from one year to another years. There are a variety

of approaches for dealing with this, all of which are beyond the scope of the introductory course. But what is definitely in the scope of the course is to point out the limitation of linear models.

The t-test. It may seem odd to discuss multiple regression and ANOVA in a course before the t-test has been covered. It's important to realize that the t-test is a special case of regression. We could, if we wanted to, carry out a t-test easily enough. Here's an equal-variance, two-sample t-test of whether swim times differ for the two sexes.

```
> t.test( time ~ sex, data=swim, var.equal=TRUE)
        Two Sample t-test
data:  time by sex
t = 5.3621, df = 52, p-value = 1.917e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.432084 16.321249
sample estimates:
mean in group F mean in group M
       66.85852        54.98185
```

In ISM, we don't use such t-test software. We prefer to deal with a general-purpose framework, not specific tests. Here is the equivalent to the t-test in terms of modeling and ANOVA.

```
> mod2 = lm( time ~ sex, data=swim)
> summary(aov(mod2))
            Df Sum Sq Mean Sq F value    Pr(>F)
sex          1 1904.2  1904.2  28.752 1.917e-06
Residuals   52 3444.0    66.2
```

Note that the p-value on the `sex` term of the model is exactly that due to the t-test.

It can be objected that with such an approach, we will leave out the unequal variance t-test. There are three good responses to such an objection. First, by using a broader technique, ANOVA, we give students the ability to generalize. It's not much of a stretch to move from two groups to three or more groups.

Second, use of the unequal variance t-test doesn't increase power by very much. If our objective is to give students a robust technique for dealing with unequal variances or non-normal distributions, let's take that objective straight on in a general way: perform the ANOVA on ranks:

```
> mod3 = lm( rank(time) ~ sex, data=swim)
> summary(aov(mod3))
            Df Sum Sq Mean Sq F value    Pr(>F)
sex          1 5420.0  5420.0  36.615 1.599e-07 ***
Residuals   52 7697.5   148.0
```

As a practical matter, we look for small p-values and our interest in using the unequal variance t-test would be to make sure that the small p-value isn't due just to unequal variances. This rank test establishes that perfectly well and is more general.

Third, and most important, any form of t-test is inappropriate because so much of the variation in swim record times is due to improvement over the years. With regression, or ANCOVA if you prefer to call it that, we can include `year` explicitly in the model. It is not a service to students in an introductory course to quibble about the second digit in a p-value, when factors that influence the first digit are being ignored!

4 The Geometry of Models

We worry about the computer becoming a black box and our students becoming mere interpreters of computer output. Students need to be able to think creatively about models and understand procedures, pitfalls, and paradoxes.

One way to think about thinking is as a way of manipulating representations. We want to choose our representations carefully; each has advantages and disadvantages.

Some dominant representations used in introductory statistics are:

- scalar statistics — single numbers such a a sample mean or standard deviation — and operations such as division and square roots.
- algebraic formulations where a symbol stands for a scalar statistic and which allow us to generalize operations.
- scatter plots that display the relationship between two variables by showing each case as an individual point.

A strong advantage of these representations is that they are familiar; almost all college-bound high-school students have encountered them, teachers and professors have mastered them. Indeed, for most people, these representations are the substance of statistics; most people see no alternative.

The disadvantages are that most students are not very good at algebra or algebraic notation; the representations obscure some relationships that are actually simple; it's difficult to generalize the representations to multiple (> 2) variables without introducting another representation, matrices, and various unfamiliar operations such as inversion.

To illustrate statistical relationships that are simple but obscured by the dominating representations, consider these "difficult" topics: orthogonality and balance in experimental design, Simpson's "paradox," multiple regression, analysis of variance and of covariance. These topics are minimized in a conventional introductory course because the representations used do not support them.

Even the correlation coefficient, r does not have a simple formulation. How many students understand this formula?

$$r = \sqrt{\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

The formula doesn't reveal much except to someone who already knows what it says. Instead of understanding what the correlation coefficient can and cannot describe about a relationship, students are reduced to treating r as a kind of speedometer: $r = 0$ indicates "no" correlation, $|r| = 1$ indicates "perfect" correlation. Unfortunately, it's the in-between situations that matter. Most troubling, the emphasis on this interpretation of r means that students are mislead into thinking that there is a single relationship between any two given variables, while in reality that relationship can depend strongly on other variables. The algebraic representation suppresses this as does the common presentation of r in terms of the rotundit of clouds on a scatter plot.

In ISM we use a representation that is rooted in geometry. Because the representation is unfamiliar, we have to teach it to them explicitly. Consider the following very small data set excerpted from some climate data for Saint Paul, Minnesota. The average monthly temperature and precipitation is shown for two months, February (2) and April (4).

Case	Month	Temperature (degrees C)	Precipitation (inches)
1	2	-6.6	1.0
2	4	7.8	2.6

In the familiar scatter plot format, these two cases would be two points on a graph. We can plot two variables at a time, say, Temperature on the x-axis and Precipitation on the y-axis, as in Figure 6.

Another format for plotting reverses the roles of the cases and variables. We call this the "case plot." Each case is be one axis, each variable will be one point. The case plot looks like this.

In a scatter plot, we can have a large number of cases; one point for each case. In a case plot, we can have a large number of variables; one point — which we will often draw as an arrow or vector — for each variable.

The limitation of a case plot is that on ordinary axes on paper it can handle only two cases, just as a scatter plot can handle only two variables. We can easily conceptualize a case plot with $N = 3$ cases: each variable will be a point in the x, y, z-coordinates of three-dimensional space. Formally, we can even think about $N = 4$-dimensional case plots or even $N = 500$-dimensional case plots. Admittedly, this is a bizarre notion to most people, and it's impossible for us to visualize a four- or 500-dimensional plot. Even a three-dimensional plot is difficult for most people. This is a significant disadvantage of this style of plotting data and is an obvious explanation for why no one would do this naturally.

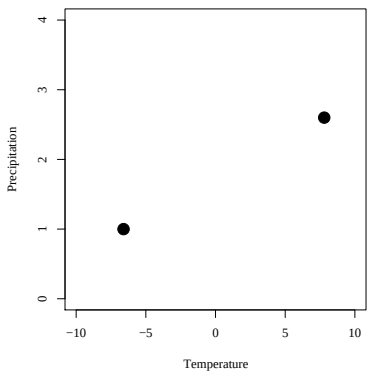


Figure 6: A scatter plot of precipitation vs temperature for $N = 2$ cases.

The case plot is definitely an unnatural approach. But, like riding a bicycle, it can be taught so that it become familiar and so that we forget how unnatural it is. The question isn't whether it is natural, but whether the advantages compensate for the need to teach something that is unnatural. We invest some days, scrape some skin, and overcome some fears as children learning to ride bicycles so that we can have a lightweight, inexpensive, maneuverable vehicle. Similarly, the case plot approach, unnatural at first, will give us clear insights into statistical processes.

Fortunately it turns out that many important statistical operations require an explanation in just two or three dimensions; by getting students to visualize the statistical operations using this geometry, they can understand what is going on when there are more than three dimensions. The low-dimensional geometry is effectively a visual metaphor for what is happening in the full-dimensional case-space.

The Basic Geometry of Statistical Modeling. The basic operations of statistical modeling are: projection, finding residuals, and measuring an angle. The basic relationship is the familiar pythagorean theorem of right triangles.

Projection. We have already seen how to use modeling notation and software to fit models. Let's look at a very simple model that says precipitation is proportional to temperature: `precipitation ~ temperature` Here is the "black-box" report from R:

```
> temperature = c(-6.6, 7.8)
> precipitation = c(1, 2.6)
> lm( precipitation ~ 1 )
```

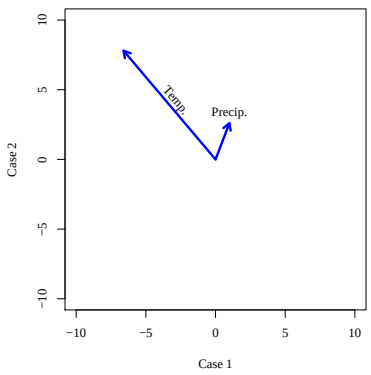


Figure 7: The case plot with two variables corresponding to Fig. 6.

```
Coefficients:
(Intercept)
        1.8
```

In case plot format we have two "variables," the genuine variable of temperature and a vector of all ones that reflects the `1` in the modeling notation, as shown in Figure 8.

The light line that runs along the ones vector is the subspace of the ones vector: all the points we can get to by taking various numbers of steps along the vector. For example, if we take 3 steps along ones, we get to the point $(3, 3)$. If we take -2.5 steps along ones, we get to the point $(-2.5, -2.5)$. The subspace is the set of all of the model points that are consistent with a model of the form `~1`

Fitting a model consists of projecting the vector of the response variable onto the model subspace. Projection means finding the point in the model subspace that is as close as possible to the response variable. Do this now. Run a pencil along the ones subspace until you reach a point that is as close as possible to the point marked precip. This point will be the best fitting model for `precipitation ~ 1`

Of course the best fitting model point is typically not the response variable point itself. The vector that joins the model point to the response variable point is the "residual vector." It's convenient to draw that vector not starting from the origin but starting at the model point. If you do this, the picture should look like Figure 9.

The fitted coefficient is the number of steps of the model vector that are needed to reach the fitted model vector. From the picture, you can see that this is a little less than two steps. In fact it is 1.8 steps; it will be exactly the

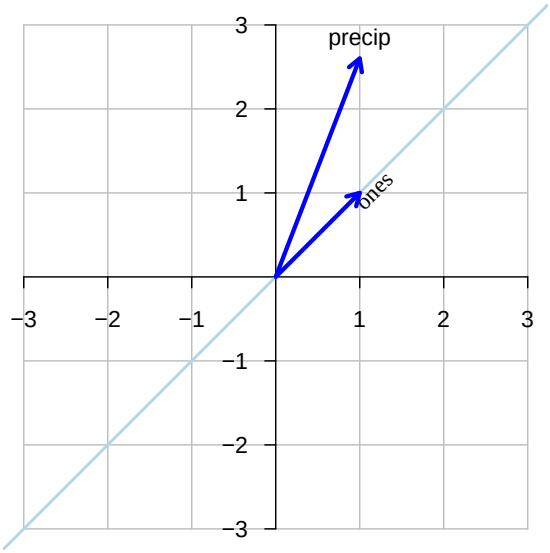


Figure 8: The case-plot setup for the simple model `precipitation ~ 1`.

coefficient found the software.

Notice that the residual vector is perpendicular to the fitted model vector. This will always be the case because we pick the fitted model vector to be the point on the model subspace that is as close as possible to the response vector.

The fundamental geometrical relationship of statistical modeling is that of the right triangle:

- fitted model vector + residual vector = response vector
- The fitted model vector and the residual vector are the legs of a right triangle. The response vector is the hypothenuse.

A basic fact you should know about vectors is that the length of the vector can be found by summing up the squares of all the coordinates and taking a square root of the sum. It's more convenient to talk about the square-length, which is the sum of squares of the coordinates. So, the vector $(1, 2.6)$ (corresponding to precipitation) has square-length $1^2 + 2.6^2 = 7.76$. (Since precipitation is measured in inches, the square length has units of inches-squared.)

Keeping in mind the right-triangle modeling relationship, it may be easy to understand that the sum of squares of the response variable equals the sum of squares of the fitted model vector added to the sum of squares of the residual vector. This relationship is fundamental to analysis of variance, which is why ANOVA tables feature sums of squares.

A t-test, geometrically. To show the power of the geometrical presentation, let's perform a t-test in both the conventional algebraic way and the geometrical

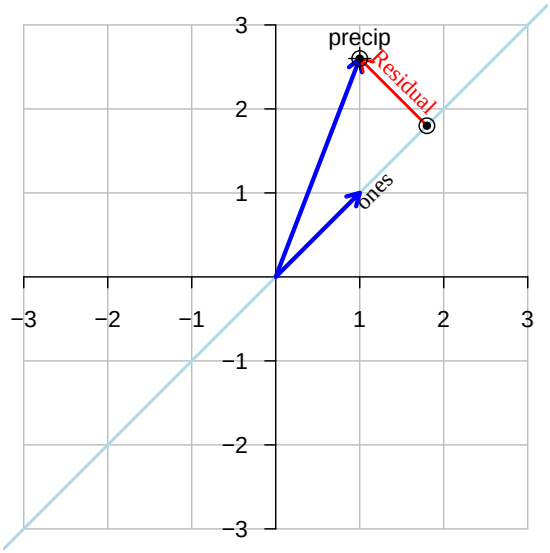


Figure 9: Projecting "precip" onto "ones" leaves a residual, which is the vector that connects the fitted model vector to the response variable vector.

way. The null hypothesis of the t-test is that the samples are drawn from a population with zero mean and unknown standard deviation. The point of a one sample t-test is to test whether the values in a sample have a mean that is inconsistent with such a zero-mean population.

As with all hypothesis testing, we first calculate a test statistic from our data and then compute a p-value from that test statistic under the assumption that the null hypothesis is true.

In the conventional approach, the test statistic is a t-value that we calculate from the mean m, standard deviation s, and number of cases N of the sample:

$$t = \frac{m}{s/\sqrt{N}} = \frac{1.8}{1.13/\sqrt{2}} = 2.25.$$

The p-value is found by looking up the test statistic in a table of the t-distribution with $N - 1$ degree of freedom. That lookup is itself non-trivial. In this case it results in $p = 0.2662$ (two sided).

If it's not obvious to you what the formula for t has to due with the null hypothesis, you are in a position similar to most students.

In the geometrical approach, the equivalent null hypothesis is that the model vector is just a random vector with no particular relationship to the response variable. In a good model, the fitted model vector is closely aligned with the response variable. A sensible test statistic to capture this is the angle between the model and the response vector. If the model vector is random, we would be surprised if it were closely aligned with the response. A very small p-value

indicates that this angle is surprisingly small. If you have a protractor, measure the angle from the figure. Or, just guess the angle by eye.

To high precision, the angle is 23.9625 degrees. (There is a simple formula for the angle in terms of addition, multiplication, and square roots that we teach to Applied Calculus and ISM students so that they can find the angle between any pair of vectors, regardless of dimension.)

What's the probability that a random model vector would be closer to the response vector than 23.9625 degrees? Because of the various symmetries (you could be on either side of the response vector, you could point positively or negatively), the answer is 23.9625 divided by 90 degrees. This gives a p-value $23.9625/90 = 2.662$.

This is not a coincidence. The t-distribution is in fact related to the distribution of random angles. The degree of freedom relates to whether the angle is in a plane, in 3-dimensional space, or so on.

Because a t-test is so simple in the geometrical approach, we don't dwell on it in ISM. Instead, we move on to the more general framework of ANOVA and ANCOVA of which the t-test and paired t-tests are special cases.

The correlation coefficient. The angle between two vectors is an intuitive measure of how closely the vectors are aligned. An angle of zero degrees means the vectors are perfectly aligned so that one vector lies in the subspace of the other. Modeling one vector by the other would give no residual: a perfect fit. Similarly, an angle of 180 degrees means the vectors are directly opposite each other so that their subspaces again overlap: a perfect fit. An angle of 90 degrees means that the projection of one vector onto the other results in a coefficient of zero: the residual from the model is as big as the response vector itself.

The standard correlation coefficient r between two variables is the cosine of the angle between the vectors, once the mean has been subtracted from each vector. The familiar R^2 statistic describing a model is simply the cosine-squared of the angle between the response variable vector and the fitted model vector (after the means have been subtracted out).

Multiple Regression More typically, we are interested in multiple explanatory terms. For example, for the model `precipitation ~ 1 + temperature` Here is the software report:

```
> lm( precipitation ~ 1 + temperature )

Coefficients:
(Intercept)  temperature
     1.7333       0.1111
```

The geometrical figure corresponding to this model involves two model vectors, as in Figure 10.

In fitting the model, we will go on a walk. First, take a choosen number of steps along the ones vector in Figure 10. Then, from the point that you

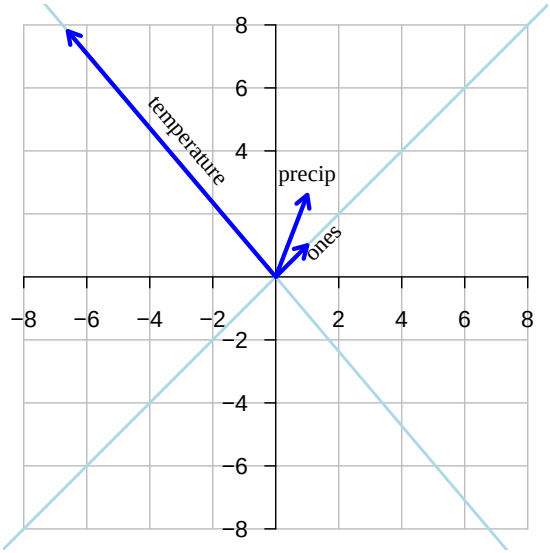


Figure 10: In multiple regression, the response is written as a linear combination of multiple model vectors.

have reached, turn in the direction of the temperature vector and take steps in that direction. Such a walk is called a "linear combination" of the two vectors. Any point you can reach in this manner is a candidate model of the form `1+temperature` The complete set of such candidates is called the "subspace spanned by the model vectors." The fitted model point is the closest of the candidates to the response variable vector.

Find the best fitting candidate now with a pencil. It takes a bit of practice. An effective procedure is to step along one of the vectors until you reach a point that is a straight shot in terms of the other vector to the response variable.

If you do this carefully, you'll find that you can get all the way to the response variable vector with the two model vectors provided. The answer is to take about two steps along "ones," then to take a very small step along "temperature." In fact, as the modeling software has already told us, the best fitted point will be 1.7333 steps along "ones" and 0.1111 steps along "temperature."

It may occur to you that we could have reached any point in the plane by taking a linear combination of "ones" and temperature. The subspace spanned by "ones" and temperature is the whole plane. This means that the best fitted model will be exact: the residual will have zero length. This shouldn't be a surprise; it's equivalent to the well known statment that there is a a straight line between connecting two points.

Higher-dimensional spaces. All of these procedures — finding angles, projecting, finding linear combinations to reach a given target point — can be done

in $N > 3$-dimensional space. Since we have trouble visualizing such spaces, we let the software do the work for us.

But for understanding what's going on, thinking about things in two- or three-dimensional space often suffices and is accessible to most people. The reason for this is that much of the "action" takes place in two- and three-dimensional subspaces of the full N-dimensional space.

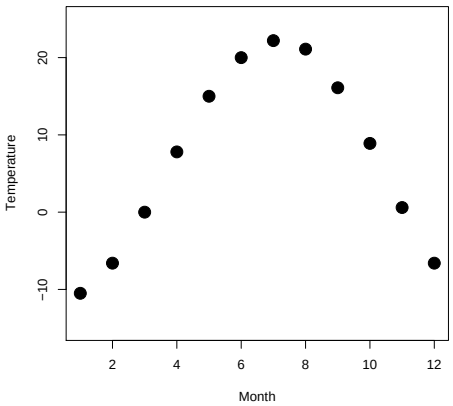
For example, consider projecting an N-dimensional response variable vector onto an N-dimesional model vector. Although the two vectors live in an N-dimensional space, the two vectors themselves span a plane: all three vectors — model, response, residual — of the model triangle will be in a plane.

When we have two different model terms and a response, we will have three N-dimensional vectors. Those three vectors live in an N-dimensional space but they span a three-dimensional space. In thinking about multiple model terms, we let the software do the work for us but we can often think about the situation by letting a single vector stand in mentally for a set of model terms.

The situation becomes a little complicated when one tries to think of the dimensionality of the space spanned by multiple model vectors. There can be redundancies — one or more of the vectors might live in the subspace spanned by other model vectors. This dimensionality, referred to as "degrees of freedom," usually follows a very simple pattern which it's easy to understand.

An r paradox ... resolved. In the previous sections, we used a ridiculously small data set with $N = 2$ cases. A critic might fairly point out that this prevents the scatter plot formalism from showing its advantages.

Here is a scatter plot of the Saint Paul, MN temperature data against month for all 12 months. We're going to see whether there is a relationship between temperature and month. The scatter plot has one point for each of the $N = 12$ cases and shows a clear seasonal relationship:



It wouldn't be unreasonable for a student to think about characterizing the

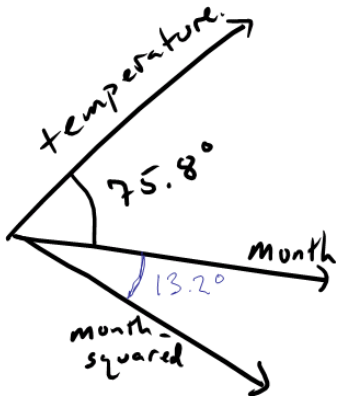
relationship between month and temperature using a correlation coefficient. For these data, that coefficient will be $r^2 = 0.06$, practically zero. (The corresponding p-value is 0.43.)

Another student comes along, notices the shape of the scatter plot is like a parabola, and suggests that temperature is related to month-squared. This student also gets a disappointment, since $r^2 = 0.0008$ — no relationship.

An ISM student, having learned that you can get more places with two vectors rather than one, tries a model with both month and month-squared. She finds for her model $R^2 = 0.95$ — a substantial relationship.

How can this be? Each of two terms individually has no relationship with the response, but taken together the two terms have an almost perfect relationship with the response. This paradox is bound to cause confusion and frustration if students have no way to think about it constructively.

Geometrically, the situation isn't hard to understand. the r^2 of 0.06 corresponds to an angle of 75.8 degrees between month and temperature. The r^2 of 0.0008 gives an angle of 88.4 degrees between month-squared and temperature. Month and month-squared are themselves correlated, an angle of 13.2 degrees. These three angles suggest the following picture:



Each of the two explanatory vectors is roughly perpendicular to the response. But if the response lies in the plane spanned by the two vectors, and apparently it does, we can get a large R^2 for the linear combination.

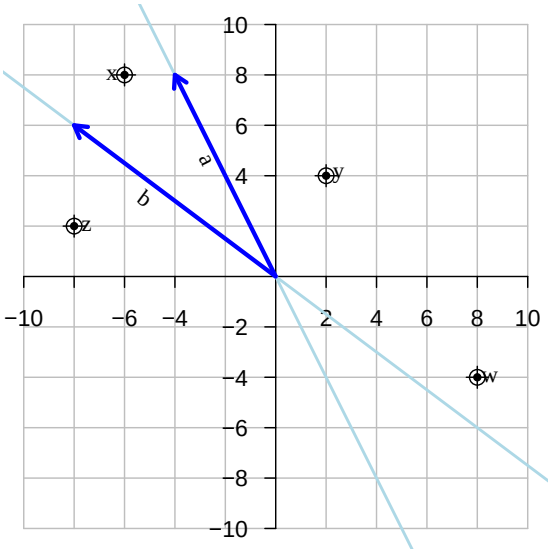
Simpson's paradox A lovely object lesson in statistics is provided by "Simpson's Paradox." The classic example of this paradox comes from a real-life situation at the University of California, Berkeley. The university was sued for sex discrimination in graduate studies because the admissions rates for women were substantially lower than those for men. The university was able to repel the suit by showing that in each department, the admissions rate for women was at least as high as for men. The reason for the conflicting patterns: women tended to apply to more competitive departments than men, so the average admissions ratio for women was less than for men.

There are many other examples. In one study, for instance, it was found that smokers had a lower death rate than non-smokers. But, it also happens that, in the population involved in the study, smokers tended to be younger than non-smokers. Even though smoking contributes strongly to the risk of death, increased age is even more potent.

In the context of models, Simpson's paradox occurs when we try to model a response y using two (or more) explanatory terms, a and b. Suppose that `y ~ a` gives a positive coefficient on x. Similarly suppose that `y ~ b` gives a positive coefficient on b. It seems intuitive that in the model `y ~ a + b` the coefficients on a and b will remain positive. Simpson's paradox is occuring when we find that one of the coefficients changes sign in the model `y ~ a + b` compared to the simpler models.

The word "paradox" emphasizes the unexpected nature of the sign reversal. When I started teaching statistics in the mid-1990s, I often used examples of Simpson's paradox because I thought they would encourage students to think about the overall situation and not to rely on simplistic descriptions. I was wrong. The dominant response to hearing about Simpson's paradox is, "I guess you can use statistics to show anything you want." It was largely in response to this typical reaction from students that I started to develop ISM. One of my goals was to remove the word "paradox" from the situation, to make it clear why the coefficients can change signs and to provide a way to deal with it so that it can be possible to have a straightforward interpretation of data.

The geometry of Simpson's paradox is shown in the figure.



The explanatory terms, a and b are correlated with one another; the angle is much less than 90 degrees. Four different hypothetical target points are shown in the figure; let's focus here on x, y, and z as representing different possible

response variables. For each of x, y, and z the coefficient on explanatory vector a, taken alone, is positive. The same for explanatory vector b. But when a and b are combined, there is a Simpson's paradox situation for response variable y and for z: these lie outside the cone defined by a and b. To see why the sign reversal occurs for y and z but not x, use a pencil and find the relevant linear combinations of a and b. For instance, to reach y: you step forward in direction a and then backward in direction b.

Simpson's paradox provides an important object lesson for scientists. If we want to avoid the ambiguity of interpretation that arises in Simpson's paradox, we should attempt, when setting up our experiments, to make our explanatory factors orthogonal to one another: at right angles. If we want to prevent the possibility of Simpson's paradox coming up in the future, based on some as-yet-unknown variable, we want to make our explanatory factors orthogonal to that as-yet-unknown variable. We can do this by assigning our treatments randomly.

Analysis of variance. If you read a conventional statistics book, you will not see any reason to believe that analysis of variance (ANOVA) is related in any way to regression. Many people, even professional statisticians, believe that analysis of variance is about looking at differences between groups rather than a general technique for drawing inferences from data. The algorithms presented for carrying out ANOVA involve many steps, the formation of mysterious intermediate quantities (sums of squares! mean squares! F ratios!), that even if the computations are done by hand they remain a black box. Interpretation of ANOVA results is difficult because the results depend on the order in which terms are specified. Then there is "analysis of covariance."

Analysis of variance is not really about differences in group means. It is about divying up credit. Consider the model depicted in Figure 11. It shows a goal point y which we want to reach with two model terms, marked a and b.

It's not important that we are looking at a $N = 2$-dimensional case plot. However many cases there are, y is one vector, b is another vector, and a is a third vector. Even if the three vectors live in a space of high dimension N, their mutual relationships can be understood entirely within the 3-dimensional space that they span. So, let us imagine that the point y is not in the plane, but is floating two units directly above the point marked in the plane.

Using just the a vector we can't get all the way to y. The same is true using just the b vector. But using both a and b we can get all the way to the point marked y in the plane. That is, the fitted model vector is the arrow reaching to the point marked y in the plane.

Since no linear combination of b and a can move us off the plane, there will still be a residual: the two units that the actual y is floating over the plane.

A quick way to characterize how effective b and a are in reaching the actual y is to compare the size of the fitted model vector to the size of the residual. A ratio will do the job nicely: the length of the fitted model vector divided by the length of the residual. As it happens, this ratio will be the tangent of the angle between the fitted model vector and the response variable vector, and so

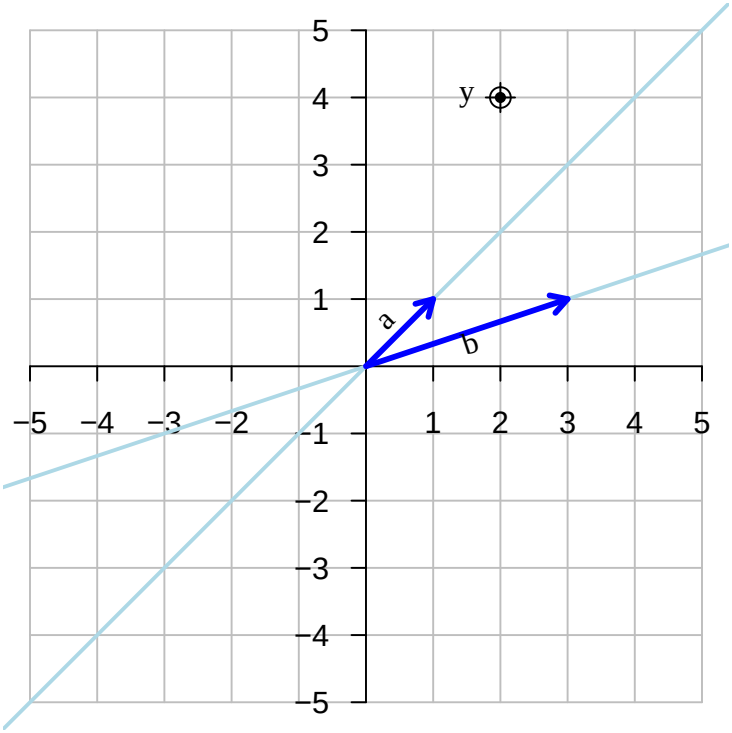


Figure 11: The geometrical set up for ANOVA of y modeled by a and b.

is related to the R^2 coefficient.)

Out of respect for the right-triangle relationship between the response variable vector (the hypothenuse) and the residual and fitted model vectors (the legs of the triangle), let's agree to use lengths-squared rather than lengths. In that case we'll look at the ratio of the lengths-squared of the fitted model vector to the residual vector; the ratio is now the tangent-squared of the angle between the two vectors and thus is still related to R^2.

A stronger reason to use lengths-squared reflects the nature of random walks. Each step of a random walk contributes equally, on average, to the mean square displacement of the overall work. But if we work not with lengths-squared but with lengths themselves (a root-mean-square displacement) we will find that early steps contribute more on average than later steps. In the physical world, this is related to the fact that a diffusing particle moves rapidly over short distances but slowly over long distances.

ANOVA is a simple accounting of how far each model terms gets us toward the response variable. We will measure the square-distance that we travel along any given model term — that will be the sum of squares for that model term.

The mean square accounts for the fact that some model terms are themselves

composed of multiple vectors, for instance the indicator vectors that are constructed from categorical variables. The number of vectors in a term is called the "degrees of freedom" of that term. In case the vectors from different model terms overlap, we need to take this into consideration by reducing the degrees of freedom. We won't go into that here except to note that the vector of all 1s always lives in the space spanned by the full set of indicator vectors of a categorical variables. That's why, when there are k different levels of categories, the degree of freedom of a categorical variable is typically $k - 1$.

Look again at the vectors above and trace the route that you would take to get to y using a and b. You would first follow direction a to the far upper right corner of the plot — five steps altogether. Then you would take a single negative step in the direction of b to reach the point marked y in the plane.

Altogether, our little journey along a and b involved a square-distance of $5^2 + 5^2 = 50$ along a and a square distance of $3^2 + 1^2 = 10$ along b. Total square-distance is 60. But this is longer than the straight line square-distance to y, which is $2^2 + 4^2 = 20$. What's happening here is that b is making up for excess distance travelled along a; we had to go past y on a in order to get back to y along direction b.

ANOVA is arranged not to give model terms ebtra credit for such shenanigans. Here is how ANOVA apportions the credit to each model term. There is a total square distance budget of 20, which is the square-distance we would travel along a straight line to get to the point best point that a and b will bring us to. We'll split up this budget between the two model terms, but we won't exceed the budget when we give credit to each model term.

To assign credit, we figure out how far a would have gotten us *if we used it on its own, without b.* You can read this off the graph; the closest point to y on a is the point $(3, 3)$, which gives a square-distance of $3^2 + 3^2 = 18$. The remainder of the budget, $20 - 18 = 2$, is assigned to b.

This seems unfair to b. If we had used model term b first, we would have found ourselves walking on b to the point $3, 1$ (which is the closest point on the subspace of b to the target y). This would give b a square-distance of $3^2 + 1^2 = 10$. So b would get 10 units of credit and a would get the remaining $20 - 10 = 10$ units.

Here are the ANOVA reports for this situation. There are two reports: one with a first and the other with b first.

```
> y = c(2,4,2)
> a = c(1,1,0)
> b = c(3,1,0)
> summary( aov( lm( y ~ a + b - 1) ) )
          Df Sum Sq Mean Sq F value Pr(>F)
a          1     18      18     4.5 0.2804
b          1      2       2     0.5 0.6082
Residuals  1      4       4
> summary( aov( lm( y ~ b + a - 1) ) )
          Df Sum Sq Mean Sq F value Pr(>F)
```

```
b          1     10      10     2.5  0.359
a          1     10      10     2.5  0.359
Residuals  1      4       4
```

The "sum of squares" reported in ANOVA is just the square-length of the contribution of each model term, calculated in the budget-respecting manner described above. The mean-square is just the square-length divided by the degrees of freedom. This is the square-distance divided by the effective number of vectors in the model term. Think of the mean square as "miles per gallon," where miles is the distance travelled and gallons tell what resources were used to get you there. (But, due to the counter-intuitive nature of random walks, with their $\sqrt{N}$ dependence, the correct measure of efficiency is miles-squared per gallon.)

The F ratio compares the mean square of a model term to the mean square of the residuals. Think of this as describing the efficiency of a model term as compared to the efficiency of a random vector living in the unexplained (that is the residual) part of the case space. The p-value is the probability of seeing such a large F value in the situation where the model terms were really just random vectors.

ANOVA is a rich way of describing how model terms can compete or collaborate in providing an explanation of the response variable. Interpreting ANOVA is complicated because ANOVA captures many of the complexities that arise when we try to divy up credit among competing or cooperating entities. If you like, think of ANOVA as describing a political process.

In the example above, the situation is "first-come, first-served." Whichever variable comes first gets more credit. But it's also possible to have situations in ANOVA, just like in any political situation, where the first player gets nothing done unless the second player is around: "last one gets the credit." A very important situation is when adding more variables reduces the mean-square of the residual; this lets even variables that make a small contribution look better. An impressive sounding name for this situation is "analysis of covariance." (Of course, adding more variables will generally reduce the residual, since they provide more routes to get closer to the response variable. But since such variables eat up degrees of freedom, they may not reduce the mean-square of the residual. This is why we study miles/gallon rather than just miles.)

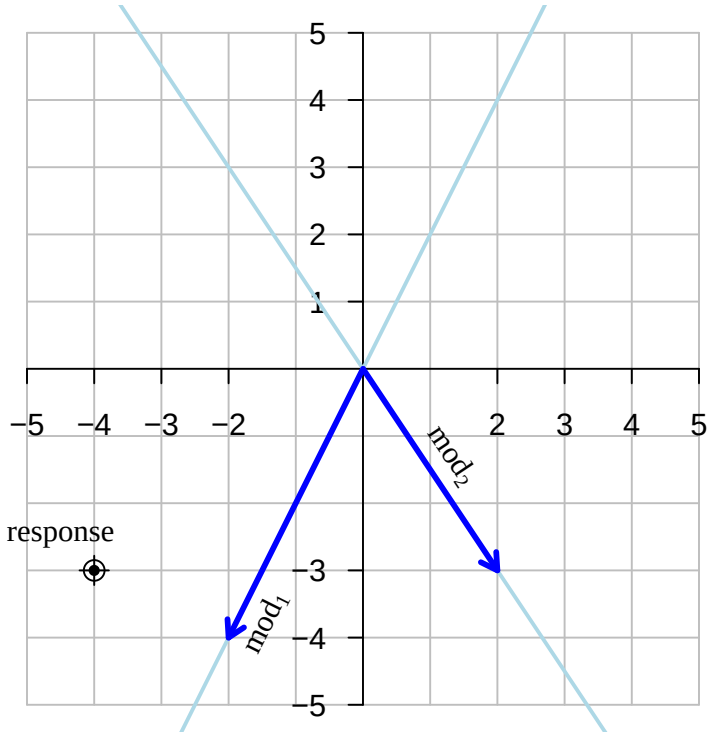
Like politics, ANOVA is complicated and requires subtlety of interpretation. It's a framework for thinking creatively about how multiple variables contribute in competing and cooperating ways. It's a shame that such a useful technique is taught to students in the extremely limiting context of a hypothesis test about group means.

Sensitivity and colinearity. Imagine two explanatory model vectors that point in almost the same direction. These two vectors still span a plane; we can find a linear combination that will bring us to any point in that plane. But what if we change one of the vectors very slightly? The linear combination that gets us to a given point may change very substantially. This is "sensitivity" or

"ill-conditioning." You can see this in Galton's data if you try to fit a model that includes three highly colinear terms: father, mother, and the interaction of father and mother. The result of the sensitivity is that the standard error on each of the coefficients becomes huge, so that each individual coefficient is no longer significant. In building models, it's important to be aware of the ramifications of colinearity. By proper experimental design we can mitigate the problem.

Exercise 4

Use a ruler and the pythagorean theorem to make an ANOVA table for the vectors shown in the figure, where response ~ mod1 + mod2. Assume that the response point is 5 units above the plane.



Exercise 5

Using the Galton data, construct several ANOVA tables involving the variables father, mother, sex, and family in various orders. Interpret each of the tables, explaining in everyday terms why the mean-square differs depending on order.

Exercise 6

In Exercise 2, part 3, we imagined a situation where the affect of smoking by a pregnant mother is to shorten gestation time, with the shortened gestation leading to decreased birth weight. Explain how this pattern would appear in an ANOVA analysis, when both smoking and gestation length are included as variables to explain birth weight, but in different orders. Contrast the pattern in the ANOVA analysis to that that would exist if smoking had a direct effect on birth weight aside from the length of gestation. [This problem doesn't involve analysing the data, just thinking about how things would look if our assumptions about mechanisms were true.]

5 Simulation and Resampling

The algorithmic approach taken by a conventional statistics course is illustrated by the way the confidence intervals are constructed. For a sample mean the steps taught to students are something like this:

1. Compute the sample mean m and sample standard deviation s from the N cases.
2. Look up the $\alpha = 0.05$ critical value in a t distribution with $N-1$ degrees of freedom. Call this $t^\star$.
3. Write down "$m \pm t^\star \frac{s}{\sqrt{N}}$ with 95% confidence."

None of these steps are terribly daunting, so what's the problem? First, the process doesn't generalize to other statistics. What if we wanted a confidence interval on a regression coefficient rather than the mean? Second, students have no general principle they can draw on to determine whether the answer they get is reasonable. If they make a mistake and forget the $\sqrt{N}$ or use N instead, or multiple by $\sqrt{N}$ rather than divide by it, they will get a wrong answer and will have no way to know that something has gone wrong. Third, the t distribution is mysterious to them. Where does it come from? "The table in the back of the book." Why not use the F or normal χ^2 distributions?

In ISM, we regard using $t^\star$ as a specialist's technique; one that's relevant to the special situation where N is very small. For even moderate sized N, the value 2 is a good stand in for $t^\star$. The algorithm we offer is somewhat more general:

1. Read the standard error from the regression report.
2. The margin of error is 2 times the standard error.

This algorithm may seem even more horribly black-box than the conventional algorithm. But using it saves time and mental energy for two topics that we think provide a solid understanding of confidence intervals and standard errors: diffusion and resampling.

We talk about the process of a random walk — diffusion — and the idea of a mean-square distance. The fundamental relationship — a relationship that's important in physics, biology, geology, finance, and statistics — is that in a random walk of N steps, the total mean-square distance, the sum of all the steps, is $\sqrt{N}$. The average step length is $1/\sqrt{N}$. We reinforce this with simulations, both on the computer and on the classroom floor. For example, we have the students huddle together and take steps in random directions, stopping at various values of N to measure the standard deviation of their displacement from the center. (This is must a matter of asking, "how far is the typical student from the center?")

By relating a statistical formula to a physical process, we make it easier for many students to think about the statistics — why does a standard error of an average scale as $1/\sqrt{N}$ while the standard error of a sum scales as $\sqrt{N}$? It also prepares students to understand why, in ANOVA, one talks about sums of squares and mean squares.

Resampling is a computational technology that is easy for students to understand and appreciate and helps them to understand what a sampling distribution is and how that relates to the confidence interval.

Resampling means simply taking a random sample with replacement from an existing sample. Here is a quick example:

```
> samp
[1] 1 2 3 4
> resample(samp)
[1] 2 4 1 3
> resample(samp)
[1] 4 4 4 4
> resample(samp)
[1] 3 3 4 3
> resample(samp)
[1] 2 3 4 4
> resample(samp)
[1] 3 2 2 1
> resample(samp)
[1] 2 3 3 3
```

Resampling mimics the process of taking a random sample from a population. In the case of resampling, the population is statistically identical to our sample, but infinite in extent.

By default, `resample` is arranged to generate a new sample with the same number of cases as the original. We can also, if we want, generate a larger sample:

```
> resample(samp, 10)
 [1] 3 2 4 2 2 4 3 1 2 1
```

To see the sampling distribution of the mean, just compute the mean of many resamples, each of the same size as the original.

```
> mean(samp)
[1] 2.5
> mean(resample(samp))
[1] 1.5
> mean(resample(samp))
[1] 2
> mean(resample(samp))
[1] 2.5
> mean(resample(samp))
[1] 3
```

We can automate this process. The `repeattrials` operator will repeatedly execute the statement for the specified number of times, collecting the results into an array.

```
> samps = repeattrials( mean(resample(samp)), 1000)
> samps
  [1] 3.25 2.50 2.75 2.25 3.25 2.50 2.00 1.75 3.00 1.75
 [11] 3.25 2.75 3.00 2.50 3.00 2.50 2.00 3.25 3.00 3.25
    and so on for 1000 trials altogether
[991] 1.75 3.25 2.25 2.50 2.00 3.50 1.75 2.75 1.00 3.00
```

The above are 1000 draws from the sampling distribution. We could make a histogram of this, or calculate its standard deviation

```
> sd(samps)
[1] 0.564217
```

This gives the standard error of the mean of a sample of size 4, because there were 4 cases in the original sample. We can simulate what would have happened if there had been, say, 400 cases in the original sample:

```
> samps = repeattrials( mean(resample(samp,400)), 1000)
> sd(samps)
[1] 0.05465623
```

Notice that with a sample size that is 100 times bigger, the standard error is reduced by a factor of $\sqrt{100}$.

More typically in ISM we would be interested in the sampling distribution of model coefficients. This can also be accomplished via resampling, which resamples a dataframe on a case-by-case basis.

Here is the model fitted on the original data

```
> lm( time ~ 1 + year + sex, data=swim)
Coefficients:
(Intercept)         year         sexM
   601.5320      -0.2751     -10.0530
```

and on a resampled data set

```
> lm( time ~ 1 + year + sex, data=resample(swim))
Coefficients:
(Intercept)         year         sexM
   591.1961      -0.2695     -10.8111
```

By repeating this calculation many times we can see what the sampling distribution of the estimated coefficients looks like. (We have typed the command on three lines.)

```
> samps = repeattrials(
+ lm( time ~ 1 + year + sex, data=resample(swim) )$coef,
+ 1000)
> samps
   (Intercept)       year       sexM
1     555.8940 -0.2515640 -10.349669
2     515.1771 -0.2314475  -8.729032
  and so on
999   631.9180 -0.2905140 -10.341338
1000  629.5846 -0.2898892  -8.377995
```

For the standard error of any of the coefficients, simply look at the standard deviation of the appropriate column:

```
> sd(samps)
(Intercept)        year        sexM
54.05861298  0.02744228  1.01599994
```

These compares well to the standard errors generated by normal theory, as was presented by the modeling software in Section 3.

Once students understand what a confidence interval is, it's no problem for them to replace 2 with 1.96 or $t^\star$ or whatever. Generalists can easily become specialists, but it's not so easy to go the other way.

If we want to show in this context that the standard error scales as $\sqrt{N}$, we can easily do so. Perhaps a more authentic demonstration comes from the following exercise.

We also use resampling to undertake hypothesis testing. By resampling an explanatory variable, we effectively implement the null hypothesis that that variable is unrelated to the others in the model. Here, for example, we will resample `sex` while leaving `year` and `time` in their original form.

```
> samps = repeattrials(lm( time ~ 1 + year + resample(sex),
+    data=swim)$coef, 1000)
> samps
   (Intercept)       year resample(sex)M
1     646.5094 -0.3007542    -0.00797899
2     647.7423 -0.3015222     0.49819874
and so on
999   643.9381 -0.2996343     0.83542832
1000  642.2846 -0.2991587     2.14890986
```

We can describe the sampling distribution under the null hypothesis in the usual ways.

```
> mean(samps)
   (Intercept)           year resample(sex)M
  646.62147879    -0.30082432     0.04514334
> sd(samps)
   (Intercept)           year resample(sex)M
   9.497681011    0.004869885    1.798907823
```

Notice that the `sexM` coefficient on the genuine data was about -10, many standard deviations away from the mean of the coefficient on the resampled data. Thus, the p-value is very small.

One advantage to using resampling is that we can present a general pattern for statistical operations: compute your test statistic on resampled data and examine the distribution of your results. We're free to use just about any test statistic we want. For example, in talking about models, we look at the overall R^2.

Exercise 7

Consider the simple model report

```
> summary(lm( heights ~ sex + father + mother, data=galton ) )
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.34476    2.74696   5.586 3.08e-08
sexM         5.22595    0.14401  36.289  < 2e-16
father       0.40598    0.02921  13.900  < 2e-16
mother       0.32150    0.03128  10.277  < 2e-16
```

We are interested in how the standard error (and, hence, the confidence interval on a coefficient) varies with the number of cases N.

To explore this, try the above model, but resampling with different numbers of cases. For instance, here is the report for $N = 100$ cases.

```
> summary(lm( heights ~ sex + father + mother, data=resample(galton,100) ) )
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.40121    7.53168   2.841 0.005484
sexM         4.78967    0.42066  11.386  < 2e-16
father       0.32664    0.09081   3.597 0.000511
mother       0.31413    0.09604   3.271 0.001491
```

How does the standard error vary with the sample size N? Does it increase or decrease with N? Does it depend linearly on N or as $\sqrt{N}$?

Exercise 8

We're going to construct a model that relates birth weight as a function of mother's weight, gestation length, and whether the mother currently smokes:

```
> summary(lm( wt ~ wt.1 + smoker + gestation + smoker:gestation, data=birth))
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.96177   10.68695   0.371 0.710919
wt.1                  0.11272    0.02267   4.972  7.6e-07
smokerTRUE          -65.65172   17.13156  -3.832 0.000134
gestation             0.37341    0.03644  10.246  < 2e-16
smokerTRUE:gestation  0.20722    0.06139   3.376 0.000761
```

The coefficients suggest that the fetuses of smokers gain weight faster than those of non-smokers for each additional day of gestation. (The coefficient `smokerTRUE` does *not* mean that smokers' babies are on average 66 ounces lighter than non-smokers; the coefficient reflects the average weight for babies with a hypothetical gestation duration of 0 days.)

We are interested in replicating this finding in another population. One question we need to address in setting up our study is how large it ought to be. We'll assume, as our alternative hypothesis, that our new population is just like the one represented in the data set. If so, we can use resampling to explore what would happen with a different sample size. For example, suppose we have a sample size of $N = 100$:

```
> summary(lm( wt ~ wt.1 + smoker + gestation + smoker:gestation,
+ data=resample(birth,100)))
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -51.59159   53.63361  -0.962  0.33864
wt.1                  0.17664    0.08178   2.160  0.03340
smokerTRUE           59.40485   72.08427   0.824  0.41204
gestation             0.54544    0.18708   2.915  0.00447
smokerTRUE:gestation -0.24469    0.25691  -0.952  0.34340
```

For this small sample, we fail to reject the null hypothesis that fetuses of smokers are growing faster than those of non-smokers.

By repeating the simulation many times, and seeing what fraction of the time we achieve significance on the term of interest, we can find the power of our study.

Find out how large the sample size needs to be to reach 90% power.