Chilean Journal of Statistics
Vol. 2, No. 1, April 2011, 31–62

Statistics Education
Research Paper

Teaching statistics: Some important tensions

George W. Cobb

Mount Holyoke College, Department of Mathematics and Statistics, Massachusetts, USA

(Received: 26 March 2010 · Accepted in final form: 26 October 2010)

Abstract

Thomas Kuhn's structure of scientific revolutions (see Kuhn, 1962) identified an essential tension between normal science and paradigm shifts. On a far more modest level, this article identifies several important tensions that confront teachers of statistics, urges all of us who teach to welcome an opportunity to rethink what we do, and argues, more narrowly, for replacing the traditional year-long sequence in probability and mathematical statistics with a one-semester course in theory and applications of linear models. Some of the general areas addressed include goals for our students, attitudes toward abstraction, the role of geometric thinking, and attitudes toward mathematics as tool and as aesthetic structure. These are illustrated by comparing different approaches to the proof of the Gauss-Markov theorem and derivation of sampling distributions.

Keywords: Coordinate free geometry · Gauss-Markov · Herschel-Maxwell · Linear models · Mathematical statistics · Statistics education.

Mathematics Subject Classification: Primary 97A99 · Secondary 62J05.

1. Introduction

As statisticians we are scientists, and as scientists we owe a continuing debt to Kuhn (1962, 1977) for recognizing and clarifying the role in science of the "essential tension" between "normal science" and times of "paradigm shift". The paradigm shifts that Kuhn wrote about were truly seismic, each creating a sudden "world out of joint". I don't presume to offer one of those seismic shifts in this article – at most I may put a few noses out of joint – but Kuhn's work convinces me that there can be value in regularly questioning tradition, even in much more modest ways. With Kuhn, I suggest that we can benefit by questioning normal science as part of doing normal science. In the same spirit, we can advance the cause of statistics teaching and learning by identifying and questioning unexamined assumptions about what we do, why we do it, and when we do it.

My goal in this article is to raise some questions about what I see as the "normal" way of teaching statistics. In raising these questions, I do not mean to suggest that the conventional way is mistaken, or even necessarily inferior to the alternatives I will describe. Rather, I merely suggest that it can be a useful exercise to rethink the standard ways of doing things, and I hope to persuade my readers that there are good reasons to regard

Corresponding address: George W. Cobb. 40 Harris Mountain Road, Amherst, MA 01002, USA.
Email: gcobb@mtholyoke.edu

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

this time of rapid change in our subject as an opportunity to question what we teach as deeply and broadly as possible.

At the same time and in parallel, I will be making an argument for teaching a course in linear statistical models as a first statistics course for students of mathematics. For many years now I have been teaching such a course to undergraduates at Mount Holyoke College in Massachusetts. The course requires no previous experience with probability or statistics, only one semester of calculus and one of matrix algebra. Thus it has fewer prerequisites than the usual first course in mathematical statistics, and so it can provide an earlier and more direct path into statistics. Like the course in mathematical statistics, a course in linear models can be mathematical enough in its content to justify being counted as an upper-division elective in a mathematics major. Unlike the mathematical statistics course, however, a course in linear models can also be a good vehicle for introducing ideas of data analysis and statistical modeling.

I have made a deliberate decision not to write this article mainly as a description of the linear models course that I have taught. Such an approach strikes me as unnecessarily narrow and limiting, because I don't expect that many readers will end up teaching such a course, and I would like to think that some of the ideas in this article will be of interest and possible value to colleagues who care about statistics education regardless of which courses they teach.

In the sections that follow, I structure my thinking as a sequence of choices, a sequence of tensions between pairs of extremes: about our goals when we teach statistics (Section 2), about how we use abstraction in our teaching (Section 3), about two geometries for representing data (Section 4), two attitudes toward mathematics (Section 5), and two ways to structure the theory of linear models (Sections 6 and 7). Sections 8 and 9 present, respectively, two approaches to the Gauss-Markov theorem and two approaches to sampling distributions. The article concludes with an argument for the centrality of linear models.

2. Two Kinds of Statistical Challenges

When we teach statistics, what is it that we want our students to learn? Surely the most common answer must be that we want our students to learn to analyze data, and certainly I share that goal. But for some students, particularly those with a strong interest and ability in mathematics, I suggest a complementary goal, one that in my opinion has not received enough explicit attention: We want these mathematically inclined students to learn to solve methodological problems. I call the two goals complementary because, as I shall argue in detail, there are essential tensions between the goals of helping students learn to analyze data and helping students learn to solve methodological problems. For a ready example of the tension, consider the role of simple, artificial examples. For teaching data analysis, these "toy" examples are often and deservedly regarded with contempt. But for developing an understanding of a methodological challenge, the ability to create a dialectical succession of toy examples and exploit their evolution is critical. As Einstein's former research assistant, John Kemeny used to tell his students at Dartmouth: "There are two kinds of mathematicians, those who use examples and those who use examples but won't admit it".

It is my position in this article that our profession needs (at least) two kinds of statisticians, those who analyze data using methodological solutions devised by others, and those who care more about devising solutions to methodological challenges than they care

about any one particular data set.[1] Our introductory applied course has been evolving, and continues to evolve, in ways that help students learn to analyze data. Our standard introduction for mathematics majors, the two semester sequence in probability and mathematical statistics, the course that ought to help students learn to solve methodological problems, has for half a century evolved mainly and sadly in Gould's punctuated sense, with our typical punctuation mark being a full stop. With a small number of notable exceptions, exceptions that I describe later in this article, the content of today's books on mathematical statistics is not much changed from that of the pioneering books by Hoel (1947) and Hogg and Craig (1959). Whereas the mathematical statistics course remains stuck in the curricular tar pits, a course in linear models can offer a vibrant mix of modeling and methodological challenges.

The purpose of the following two examples is to illustrate first how a linear models course lends itself to rich analyses and modeling challenges based on complex data sets, and second how the same course can also be structured as a succession of methodological challenges.

Example 2.1 [Faculty salaries] The scatter plot in Figure 1 plots mean academic salary versus the percentage of faculty that are women, with points for 28 academic subjects. The data come from a national survey of universities; the complete data set is given in Appendix 1, reproduced from Bellas and Reskin (1994).

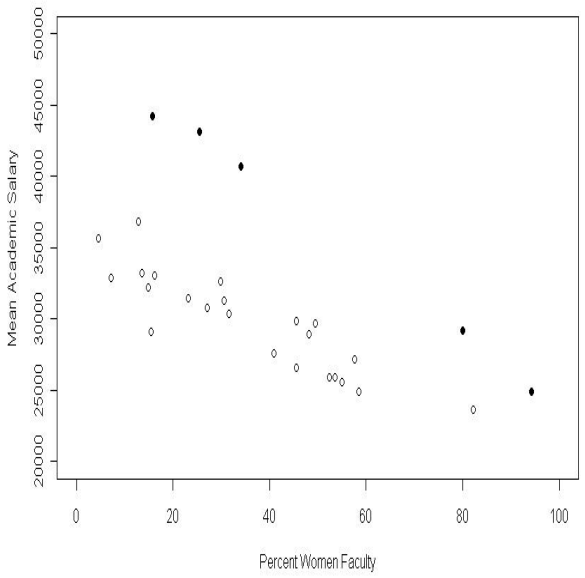

Figure 1. Mean academic salary versus percentage women faculty, for 28 academic subjects.

My decision to blacken five dots and label the axes gives away a lot. Please ignore all that as you consider the first of four sets of questions that arise naturally in the context of this data set.

Question 1. Pattern and context. Imagine seeing just the points, with no axis labels, and none of the points blackened: Based on abstract pattern alone, how many clusters do you

[1]I wrote "at least two kinds", because I salute a third valuable contributor to statistics, the abstract synthesizer. For an example, consider the work by Dempster et al. (1977), which unified decades of work and dozens of research publications by recognizing the unifying structure of what they called the "EM algorithm".

see? (Many people see three: a main downward sloping cluster of 22 points, a trio of roughly collinear outliers above the main cluster, and a triangle of points in the lower right corner.) Now rethink the question using the axis labels, and the fact that the points are academic subjects. The three highest paid subjects are dentistry, medicine, and law. The other two darkened circles are nursing and social work. All five of these subjects require a license to practice, which would tend to limit supply and raise salaries. (The remaining subject in the triangle is library science, which does not require a license.) With the additional information from the applied context, two linear clusters now seems a better summary than one main group with two sets of outliers. Adding context has changed how we see the patterns.

Question 2. Lurking variables, confounding, and cause. There is a strong negative relationship between salary and percent women faculty. Is this evidence of discrimination? Labeling the points suggests a lurking variable. Engineering and physics are at the extreme left, with the most males and the most dollars. Just to the right of engineering and physics are agriculture, economics, chemistry, and mathematics, also having few women faculty and comparatively high salaries. Music, art, journalism, and foreign languages form a cluster toward the lower right; all have more than 50% women faculty and salaries ranking in the bottom third. Nursing, social work, and library science form the triangle at the far right, where both men and money are in shortest supply. The overall pattern is strong: heavily quantitative subjects pay better, and have fewer women; subjects in the Humanities pay less, and have more women. How can we disentangle the confounding of men and mathematics?

Question 3. One slope, or two; correlation and transforming. The five darkened points lie along a line whose slope is steeper than a line fitted to the other 23 points. Is the difference in slopes worth an extra parameter? How can we measure the strength of a fitted relationship? If we convert salaries to logs, how does our measure of fit change? Does converting to logs make a one-slope model (two parallel lines) fit better?

Question 4. Adjusting for other variables; added variable plots, partial and multiple correlation. There is no easy way to measure the confounding variable directly, but the complete data set includes additional variables related to supply and demand: the unemployment rate in each subject, the percentage of non-academic jobs, and the median non-academic salary. The correlations tend to be about what you would expect: subjects in the humanities have higher unemployment, fewer non-academic jobs, and lower non-academic salaries. How can an analysis take into account these economic variables, make appropriate adjustments, and see whether the remaining pattern shows evidence of discrimination?

This data set and corresponding open-ended questions are typical of a great many that can serve as examples in either a second applied course or a first statistics course in linear models. They are not typical, however, of what we see in the probability and mathematical statistics sequence. As I hope to convince you in Example 2.3, when data sets are included in books on mathematical statistics they tend to be chosen to illustrate a single concept or method, perhaps two, and they too often lack the open-ended quality that research in statistics education encourages us to offer our students.

When I teach a course for mathematic majors, I of course want them to learn about data analysis, but I also want them to develop solutions for methodological challenges. Fortunately, teaching least squares makes it natural to combine data analysis problems and methodological questions in the same course.

Example 2.2 [The Earth, the Moon, Saturn: Inconsistent Systems of Linear Equations] The origins of least squares date back to three problems from eighteenth century science, all of which led to inconsistent sets of linear equations at a time when there was no accepted method for "solving" them; for more details, see Stigler (1990).

The shape of the Earth. Newton's theory predicts that a spinning planet will flatten at the poles. If Newton is right, then the earth should have a larger diameter at the equator than from pole to pole, with ratio 231/230. To check the prediction, scientists made observations which led, after simplification, to an inconsistent set of linear equations that had to be "solved" to answer the question.

The path of the Moon. Centuries ago, navigation depended on knowing the path of the moon, but the moon wasn't behaving as predicted. Careful observation and theory led again to an inconsistent linear system whose "solution" was needed.

Saturn and Jupiter. The motions of Saturn and Jupiter show apparent anomalies suggesting that the smaller planet might fly out of the solar system while the heavier one would slowly spiral into the sun. Understanding these anomalies, too, led to an inconsistent linear system.

Like the AAUP example, this one also leads to an open-ended set of questions, but this time the challenge is methodological: What is a good way to reconcile an inconsistent set of linear equations? One approach, used by Tobias Mayer in 1750 (see Stigler, 1990, pp. 16 ff.) is to reduce the number of equations by adding or averaging. A more sophisticated variant, used by Laplace in 1788 (see Stigler, 1990, pp. 31 ff), is to recognize natural clusters and form suitable linear combinations. An alternative approach is closer to the spirit of least squares: find a solution that minimizes the largest absolute error, or the sum of the absolute errors, or ... As students explore these possible solutions, they develop a sense of properties of a good method: it should be free from ambiguity, so that all practitioners agree on the solution; it should produce a solution; it should produce only one solution; and it should be analytically tractable.

One end result of this exploration is that students come to recognize that the least squares solution came comparatively late, after earlier approaches had been tried and found wanting. A second, deeper, end result is that students see an abstract structure to the solution: applied problems lead to an abstract methodological challenge, whose solution requires first choosing criteria for success, then using mathematics to satisfy the criteria.

A second methodological challenge in the same spirit as the challenge of solving an inconsistent linear system is to find a measure of how well the "solution" solves the system, in statistical language, to find a measure of goodness of fit. Residual sum of squares seems a natural choice, but working with simple examples reveals that it is not scale invariant. Dividing by raw sum of squares solves the problem of scale invariance, but the revised measure is no longer location invariant. Dividing instead by the mean-adjusted total sum of squares solves the problem, in a way that generalizes easily to multiple predictors. Moreover, the process of adjusting both the numerator and denominator sums of squares can lead later on to partial correlations and added variable plots.

Yet a third methodological challenge is to develop a measure of influence. Some experimentation with examples reveals that changing the value of y_i and plotting $\hat{y}_i$ versus y_i gives a linear relationship, which suggests that points with steeper slopes have greater influence, and raises the question of how to find the slope without doing the plot. At this point I typically refer the students to Hoaglin and Welsch (1978), and ask them to prove some of the results in that article.

Note that all three of these challenges can be addressed without relying on probability, which not only makes the challenges accessible to students who have no background in probability or statistics, but also makes the results applicable in applied settings where

distributional assumptions might be hard to justify. (I return to this point in Sections 6 and 7).

A fourth and final methodological challenge that can be addressed without probability is to quantify multicollinearity and its effect on model choosing and fitting. One might be tempted to conclude that the usual analysis in terms of "variance inflation factors" necessarily involves probability, but while a probabilistic interpretation can be both relevant and useful, collinearity can be addressed in a distribution-free setting, as a purely geometric and data-analytic phenomenon.

Examples 2.1 and 2.2 have illustrated how, in the context of a course on linear models, it is possible to pose both data analytic and methodological challenges. Notice that neither of the standard introductions to statistics offers the same variety of challenges. The usual first course with applied emphasis is not suited to offering methodological challenges, mainly because it is pitched at a comparatively low level mathematically. Moreover, although such courses do directly address analysis of data, they don't ordinarily begin with an open-ended data-analytic challenge that will eventually call for multiple regression methods, as in Example 2.1. The methods taught in a typical applied first course – e.g., t-tests, tests for proportions, simple linear regression – do no lend themselves to interesting modeling challenges the way a least squares course does.[2]

Of course my comparison is unfair: the introductory course is not designed for mathematically sophisticated students who have the background for a course in linear models. To be fair, then, consider what we offer students who take a course in mathematical statistics.

Although it is possible to teach the mathematical statistics course as a succession of methodological challenges (see, in particular Horton, 2010; Horton et al., 2004), the course content does not ordinarily lend itself to interesting data analytic questions in the same way that a linear models course can. (But see Nolan and Speed, 2000, for a striking, original, and valuable book that swims bravely against the current.) Within the mainstream, however, consider four pioneering books that have earned my admiration because of the way they have anchored theory in the world of real data: in probability, Breiman (1969), and Olkin et al. (1980); and in mathematical statistics, Larsen and Marx (1986), and Rice (1995). Much as I applaud these books and their authors, I nevertheless characterize their use of real data largely as "illustrative". When we teach linear models, it is easy to pose data-based questions that are open-ended ("Evaluate the evidence of possible discrimination against women in academic salaries"). When we teach probability or mathematical statistics, our questions tend to be much more narrowly focused. The following example illustrates two data-based problems from mathematical statistics courses, and two methodological challenges.

Example 2.3 [Engine Bearings, Hubble's Constant, Enemy Tanks, SD from IQR]

Engine bearings [Rice (1995, p. 427)] "A study was done to compare the performances of engine bearings made of different compounds ... Ten bearings of each type were tested. The following table gives the times until failure ... (i) Use normal theory to test the hypothesis that there is no difference between the two types of bearings. (ii) Test the same hypothesis using a non-parametric method. (iii) Which of the methods ... do you think is better in this case? (iv) Estimate π, the probability that a type I bearing will outlast a type II bearing. (v) Use the bootstrap to estimate the sampling distribution of $\hat{\pi}$ and its standard error". Comment: this is a thoughtful exercise whose multiple parts are coordinated in a way that takes the task beyond mere computational practice. All the same, this exercise does not offer the kind of modeling challenge that is possible in a course on linear models.

[2] A notable exception is the book by Kaplan (2009), which takes a modeling approach in a first course.

HUBBLE'S CONSTANT [Larsen and Marx (1986, pp. 450-453)] Hubble's Law says that a galaxy's distance from another galaxy is directly proportional to its recession velocity from that second galaxy, with the constant of proportionality equal to the age of the universe. After giving distances and velocities for 11 galactic clusters, the authors illustrate the computation of the least squares slope H for the model $v = Hd$. Comment: this example is made interesting by the data, and the fact that the reciprocal of H is an estimate for the age of the universe. However, there is no data analytic challenge: the model is given, and fits the data well.

ENEMY TANKS [Johnson (1994)] Suppose tanks are numbered consecutively by integers, and that tanks are captured independently of each other and with equal chances. Use the serial numbers of captured tanks to estimate the total number of tanks that have been produced. Abstractly, given a simple random sample $X_1, \ldots, X_n$ from a population of consecutive integers $\{1, \ldots, n\}$, find the "best" estimate for N. Comment: this problem is so simple in its structure and so removed from data analysis that it almost qualifies as a "toy" example. Nevertheless, it offers a proven effective concrete context that is well-suited to thinking about particular estimation rules, general methods for finding such rules, and criteria for evaluating estimators.

SD FROM IQR [Horton (2010)] "Assume that we observe n iid observations from a normal distribution. Questions: (i) Use the IQR of the list to estimate σ. (ii) Use simulation to assess the variability of this estimator for samples of $n = 100$ and $n = 400$. (iii) How does the variability of this estimator compare to $\hat{\sigma}$ (usual estimator)?" Comment: answering this question requires a mix of theory and simulation, and students explore important ideas and learn important facts in return for their efforts. Yet it is also typical of the way that the content of our consensus curriculum for the probability and mathematical statistics courses tends to bound us and our students away from data analysis. (I return to this point in the final section.)

As a matter of personal preference, I'm very much in sympathy with the approach of Horton (2010): I like to structure the entire course in linear models as a sequence of methodological challenges, as set out in Appendix 2. Others might prefer instead to insert just one or a few such challenges into a course, whether linear models or probability or mathematical statistics.

Of course if your main goal in a linear models course is to teach your students to analyze data, you don't want to spend a lot of time on the logic and choices that lead from questions to methods; you naturally want to focus on using those methods to learn from data. To some extent the decision about goals depends on one's attitude toward the role of abstraction in a particular course.

3. Two Attitudes Toward Abstraction

When it comes to abstraction, there is an essential tension between wholesale and retail, nicely captured by Benjamin Franklin's childhood impatience with his father's habit of saying a lengthy blessing before each meal. "Why not save time", young Ben asked, "by saying a single monthly blessing for the whole larder?" Franklin senior was not amused. He thought there was value in systematic, concrete repetition with minor variations. In this section, much as I sympathize with Franklin junior's wish for abstract efficiency, I end up siding with Franklin senior and his recognition that understanding grows from repeated encounters with concrete examples.

When it comes to teaching statistics, we teachers and authors recognize that abstract formulations can be both precise and efficient, and the conventional attitude seems often to be that our exposition should be as abstract as our students and readers can manage. On this view, the only check on abstract exposition is the ceiling imposed by the capacity of our audience.

Consider, for example, how efficient we can be if we rely on the standard template for mathematical exposition – definition, example, theorem, proof – to present the linear model.

EXAMPLE 3.1 [AN ABSTRACT EXPOSITION OF THE LINEAR MODEL] (see, e.g., Hocking, 1996, p. 20, Graybill, 1961, p. 109, and Searle, 1971, p. 79)

DEFINITION. A linear model is an equation of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ is an $n \times 1$ vector of observed response values, $\boldsymbol{X}$ is an $n \times p$ matrix of observed covariate values, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters to be estimated from the data, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of unobserved errors.[3]

ILLUSTRATION. For the AAUP data of Example 2.1, consider the model $Y_i = \beta_0 x_{i0} + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where for subject $i = 1, \ldots, 28$, Y_i is the academic salary, $x_{0i} = 1$, x_{1i} is the percent women faculty, and x_{2i} is an indicator, equal to 1 if the subject requires a license, 0 otherwise.

DEFINITION. The principle of least squares says to choose the values of the unknown parameters that minimize the sum of squared errors, namely, $Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$.

THEOREM. $Q(\boldsymbol{\beta})$ is minimized by any $\boldsymbol{\beta}$ that satisfies the normal equations $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{Y}$. If the coefficient matrix $\boldsymbol{X}^\top \boldsymbol{X}$ has an inverse, there is a unique least squares solution $\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$.

Depending on the intended readership and emphasis, an exposition as formal and compact as this may be entirely appropriate. However, for some courses, a much less common alternative approach may offer advantages. For this approach, my goal is for students to develop an abstract understanding themselves, working from simple, concrete examples, looking for patterns that generalize, eventually finding a compact formal summary, and then looking for reasons for the pattern. The normal equations lend themselves in a natural way to this approach.

EXAMPLE 3.2 [PATTERNS IN NORMAL EQUATIONS]

BACKGROUND. This exercise assumes that students have already seen applied settings for all the models that appear in the exercise, and that students who have not seen partial derivatives in a previous course have been given an explanation of how to extend the logic of finding the minimum of a quadratic function of a single variable to functions of two or more variables.

EXERCISE. Find the normal equations for the following linear models. Start by using calculus to minimize the sums of squares, but keep an eye out for patterns. Try to reach a point where you can write the set of normal equations directly from the model without doing any calculus:

[3] Distributional assumptions are treated in later sections.

(a) $Y_i = \alpha + \varepsilon_i$,
(b) $Y_i = \beta x_i + \varepsilon_i$,
(c) $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,
(d) $Y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$,
(e) $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$.

Results. Students recognize that for each of these models, the number of partial derivatives equals the number of unknown parameters, and so the linear system will have a square matrix of coefficients and a vector of right-hand-side values, giving it the form: $\boldsymbol{C\beta} = \boldsymbol{c}$. Moreover, students recognize that the individual coefficients and right-hand values are sums of products, and in fact are dot products of vectors. This leads naturally to rewriting the model in vector form $\boldsymbol{Y} = \beta_0 \boldsymbol{1} + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_k \boldsymbol{x}_k + \boldsymbol{\varepsilon}$ and stating explicitly that the coefficient in equation i of β_j is $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$, with the right-hand side $\boldsymbol{x}_i \cdot \boldsymbol{Y}$ and $\boldsymbol{1}$ being a vector of ones. From there, it is but a short step to the matrix version of the model and $\boldsymbol{X}^\top \boldsymbol{X\beta} = \boldsymbol{X}^\top \boldsymbol{Y}$.

Is there a quick way to see why the normal equations follow this pattern? One standard way is to introduce notation for vectors of partial derivatives and set the gradient to zero, but I regard this as little more than new notation for the same ideas as before.[4] An alternative that deepens understanding is based on geometry, rather than more calculus. The pattern in the normal equations has already suggested the usefulness of the vector form of the model $\boldsymbol{Y} = \beta_0 \boldsymbol{1} + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_k \boldsymbol{x}_k + \boldsymbol{\varepsilon}$. Apart from the error term, we are fitting the response $\boldsymbol{Y}$ using a linear combination of vectors, that is, by choosing a particular element of the subspace spanned by the columns of $\boldsymbol{X}$. Which one? The one that minimizes the squared distance $\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2$, namely, the perpendicular projection of $\boldsymbol{Y}$ onto the subspace. Some students may have done enough already with the geometry of $\mathbb{R}^n$ to be able to benefit from so brief an argument. Others will need to spend time developing the geometry of variable space, as in the next section.

Almost every topic we teach offers us a range of choices, from the efficient, abstract, top-down approach of Franklin junior at one extreme to Franklin senior's slower, bottom up approach based on concrete examples. This is the essential tension between traditional teaching and the method of discovery, of R.L. Moore, and of constructivism. Because the content of a course in linear models is so highly structured, while at the same time the models and applied settings are so varied, teaching a course in linear models offers the instructor an unusually rich set of possibilities for choosing between abstract exposition and teaching through discovery.

4. Two Ways to Visualize the Data: Individual Space and Variable Space

It is well-known, and long-known, that the standard picture for representing a least squares problem has a dual, one that dates back at least to Bartlett (1933). A seminal paper is Kruskal (1961), which treats the dual geometry with a "coordinate free" approach; see also Eaton (2007). Dempster (1968) labels the two complementary pictures as "individual space" and "variable space", and is explicit that the two pictures are related (with appropriate minor adjustments) in the same way that dual vector spaces are related. Bryant (1984) offers an elegant, brief, and elementary introduction to the basic geometry of variable space and its connections to probability and statistics. Herr (1980), reviews eight major articles as the core of his brief historical survey of the use of this geometry

[4] Unless the course takes the time to connect the vector of partials with the direction of steepest descent.

in statistics. Textbooks that offer a substantive treatment of this geometry include, in chronological order, Fraser (1958), Scheffe (1959), Rao (1965), Draper and Smith (1966), Dempster (1968), Box et al. (1978), Christensen (1987), Saville and Wood (1991), Kaplan (2009), and Pruim (2011), among others. Despite the existence of this list. however, on balance the geometry of variable space plays only a small part in the mainstream exposition of linear models.

In the remainder of this section, I first introduce the two complementary pictures by way of a simple example, then discuss some possible consequences for teaching linear models.

EXAMPLE 4.1 [THE CRYSTAL PROBLEM (adapted from Moore, 1992)]

Let β be the width of a one-celled crystal, and let 2β be the width of a four-celled crystal, as in Figure 2. Suppose we want to estimate β, and we have two measurements, one for each crystal: $Y_1 = \beta + \varepsilon_1$ and $Y_2 = 2\beta + \varepsilon_2$; see Figure 2. In order to keep the arithmetic and diagrams cleaner, I have assumed unrealistically large errors of 2 and -1, giving observed widths of $y_1 = 3$ for the smaller crystal and $y_2 = 1$ for the large one.

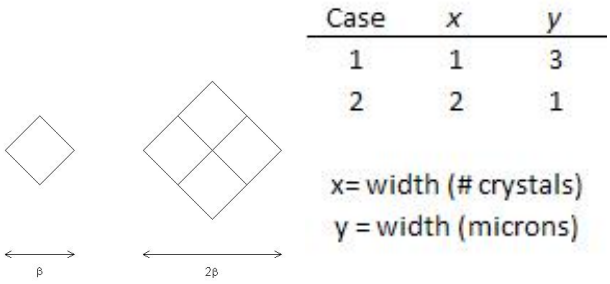

Figure 2. The crystal problem.

The usual "individual space" picture for this problem (Figure 3, left panel) is the familiar scatter plot, with each case (observation) plotted as a point (x, y) whose coordinates are determined by the variables. In this representation, the set of possible models – the set of all βx with $\beta \in \mathbb{R}$ is the pencil of lines through the origin, and the least squares principle says to choose the line that minimizes the sum of squared vertical deviations from the data points to the line.

The "variable space" representation (Figure 3, right panel) plots each variable as a vector, with cases corresponding to coordinate axes. Now the set of all possible models or "model space" – in this instance all scalar multiples $\boldsymbol{X\beta}$ – is the subspace spanned by $\boldsymbol{x}$. The sum of squared residuals $\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2$ is the squared distance from Y to $\beta\,\boldsymbol{x}$, a quantity that is minimized by taking the perpendicular projection of the response vector $\boldsymbol{Y}$ onto model space.

More generally, for any linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, model space $\langle \boldsymbol{X} \rangle$ is the column space of $\boldsymbol{X}$, that is, the set of all linear combinations of the columns of $\boldsymbol{X}$. For any choice of the parameter vector $\boldsymbol{\beta}$, the product $\boldsymbol{X\beta}$ lies in model space, and the squared length of the difference vector $\boldsymbol{Y} - \boldsymbol{X\beta}$ equals the residual sum of squares. For this general case the dimension of variable space equals the number n of cases, and the picture is impossible to draw, but a useful schematic version is possible, as shown in Figure 4.

The horizontal plane represents model space, the p-dimensional subspace of all linear combinations $\beta_0\mathbf{1} + \beta_1\boldsymbol{x}_1 + \cdots + \beta_k\boldsymbol{x}_k$ of the columns of $\boldsymbol{X}$. The vertical line represents error space $\langle \boldsymbol{X} \rangle^{\perp}$, the $(n - p)$-dimensional subspace that is the orthogonal complement of model space. The vector $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ of fitted values is the orthogonal projection of the

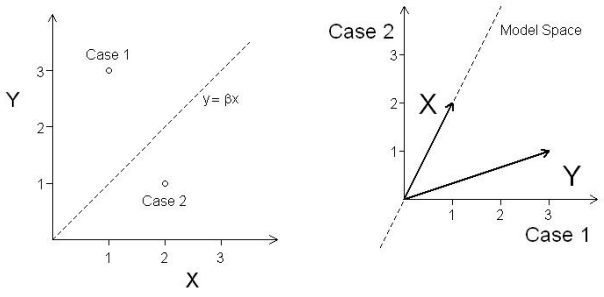

Figure 3. Individual space and variable space pictures for the crystal problem For the individual space picture on the left, each case is plotted as a point, and each coordinate axis corresponds to a variable. For the variable space picture on the right, each variable is plotted as a vector, and each case corresponds to a coordinate axis. Model space is the set of all linear combinations of the predictor variables; error space is the orthogonal complement of variable space, and the least squares estimator is the perpendicular projection of the response vector onto model space.

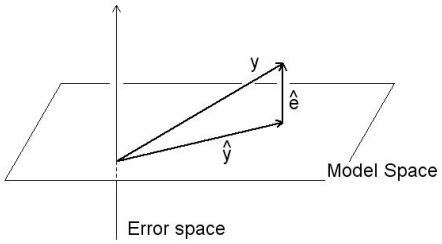

Figure 4. A schematic representation of variable space.

response vector y onto model space, and the residual vector $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is the orthogonal projection of y onto error space. The vectors $\boldsymbol{Y}$, $\hat{\boldsymbol{Y}}$, and $\hat{\varepsilon}$ form a right triangle.

Once a student has become familiar with the geometry of variable space, there is a short geometric derivation of the normal equations:

(a) The sum of squares function $Q(\boldsymbol{\beta})$ is the squared Euclidean distance from $\boldsymbol{Y}$ to $\boldsymbol{X\beta}$.
(b) That distance is minimized by $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, the orthogonal projection of $\boldsymbol{Y}$ onto model space.
(c) For the projection to be orthogonal, the difference vector must be orthogonal to model space, i.e., $\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X\beta}) = \boldsymbol{0}$.

Depending on a student's linear algebra background, it can take time to develop the geometry of variable space, time that could instead be devoted to something else. Is it worth it? The answer, of course, depends on a teacher's goals and priorities, but learning the geometric approach offers a variety of benefits.

Perhaps the greatest benefit is the directness and simplicity of the geometric argument. Compare, for example, the geometric and calculus-based derivations of the normal equations, and notice how the calculus-based argument requires one to think about the structure of the quadratic function $Q(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2$ in order to justify the use of partial derivatives to find the minimizing value of $\boldsymbol{\beta}$. In effect the calculus approach requires both the individual space picture and an auxiliary picture of the graph of Q. This extra picture takes students on a cognitive detour, because the picture is not in any sense a representation of the data, and once we have the normal equations, we have no more need for the picture. In contrast, the variable space picture simultaneously represents the data and contains within it a simple way to visualize $Q(\boldsymbol{\beta})$ as a squared distance. Not only is the minimizing $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ visually apparent, but, moreover, the orthogonality of $\boldsymbol{X}$ and $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ gives the normal equations at once. After that, the same geometric representation

can be used, over and over, in a variety of contexts, to deepen one's understanding of basic concepts and results. All correlations, whether, simple, multiple, or partial, are cosines of angles in variable space, and thus the restriction of values to [-1,1] and invariance with respect to location and scale changes are simple consequences of the geometry. Every sum of squares is a squared distance; every partitioning of a sum of squares is tied to a right triangle, and the usual F-statistic for nested models is, apart from a scaling factor, the squared cotangent of an angle:

Pythagorean relationships.

$$\|\boldsymbol{Y} - \bar{\boldsymbol{Y}}\,\mathbf{1}\|^2 + \|\bar{\boldsymbol{Y}}\,\mathbf{1}\|^2 = \|\boldsymbol{Y}\|^2$$
$$\mathrm{SS}_{\mathrm{Regression}} + \mathrm{SS}_{\mathrm{Residual}} = \mathrm{SS}_{\mathrm{Total}}$$
$$\|\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\,\mathbf{1}\|^2 + \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \|\boldsymbol{Y} - \bar{\boldsymbol{Y}}\,\mathbf{1}\|^2.$$

Correlation and angle.

(a) Simple. $\mathrm{Corr}(X, Y) = \cos(\theta_{XY.1})$ where $\theta_{XY.\mathbf{1}}$ is the angle between $\boldsymbol{X} - \bar{\boldsymbol{X}}\,\mathbf{1}$ and $\boldsymbol{Y} - \bar{\boldsymbol{Y}}\,\mathbf{1}$.
(b) Coefficient of determination. $R^2 = \cos^2(\boldsymbol{\theta}_{Y\hat{Y}.1})$ where $\boldsymbol{\theta}_{Y\hat{Y}.1}$ is the angle between $\boldsymbol{Y} - \bar{\boldsymbol{Y}}\,\mathbf{1}$ and $\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\,\mathbf{1}$.
(c) Partial correlation. $\mathrm{Corr}(X, Y|Z) = \cos(\theta_{XY.1Z})$, where $\theta_{XY.\mathbf{1}Z}$ is the angle between the two difference vectors obtained by projecting $\boldsymbol{X}$ and $\boldsymbol{Y}$ onto the subspace spanned by $\mathbf{1}$ and $\boldsymbol{Z}$.

General regression significance test (nested F-test).

$$F = \frac{\dfrac{\mathrm{RSS}_{\mathrm{Full}} - \mathrm{RSS}_{\mathrm{Reduced}}}{\mathrm{df}_{\mathrm{Reduced}} - \mathrm{df}_{\mathrm{Full}}}}{\dfrac{\mathrm{RSS}_{\mathrm{Full}}}{\mathrm{df}_{\mathrm{Full}}}} \propto \cot^2(\theta),$$

where $\mathrm{RSS}_{\mathrm{Full}}$ and $\mathrm{RSS}_{\mathrm{Reduced}}$ are residual sums of squares for two nested linear models having residual degrees of freedom $\mathrm{df}_{\mathrm{Full}}$ and $\mathrm{df}_{\mathrm{Reduced}}$, and θ is the angle between $\hat{\boldsymbol{Y}}_{\mathrm{Full}} - \bar{\boldsymbol{Y}}\,\mathbf{1}$ and $\hat{\boldsymbol{Y}}_{\mathrm{Reduced}} - \bar{\boldsymbol{Y}}\,\mathbf{1}$.

Finally, the geometry of variable space can provide a useful finite-dimensional introduction to the geometry of Hilbert space that some students will need later when they study Fourier series and stochastic processes.

Quite apart from its value for understanding statistical ideas, the geometry of variable space has an inherent aesthetic appeal for some students. As I argue in the next section, even in a course with an applied emphasis, there can be good reasons to attend to issues of mathematics for its own sake.

5. Two Attitudes Toward Mathematics: As a Tool or for its Own Sake?

When teaching subjects like physics, or economics, or statistics, it is common to regard mathematics as a tool, and thus to regard mathematics as a means to an end, not as an end in itself. In teaching the sciences, where mathematics is a means to an end, getting to the destination efficiently is a guiding principle, and the aesthetics of the path is secondary. To a pure mathematician, however, mathematics is an aesthetic object, one that Bertrand

Russell compared to sculpture because of the cold austerity of its renunciation of context. Physicists use mathematics to study matter, economists use mathematics to study money, and statisticians use mathematics to study data, but mathematicians themselves boil away the applied context, be it matter or money or data, in order to study the clear crystalline residue of pure pattern. As a former colleague once put it, mathematics is the art form for which the medium is the mind.

In this section I suggest that even though as statisticians we often and appropriately regard mathematics as a tool and put a priority on efficiency of derivation and exposition, there are reasons to regard mathematics also as an end in itself, and, at times, to sacrifice expository efficiency in order to teach in a way that celebrates mathematics as an aesthetic structure. Before presenting an argument, however, it seems useful to be more concrete about what I mean by the aesthetic aspects of mathematics, and how teaching with mathematics as an end in itself differs from teaching with mathematics as a means to an end.

I don't have anywhere near the qualifications (or for that matter, the patience) to attempt an aesthetic analysis that would be worthy of a philosopher. Instead, I shall focus on just one important feature that to me helps distinguish the mathematical aesthetic, namely, surprise connections revealed by abstract understanding. Over and over in mathematics, things that seem completely different on the surface turn out, when understood at their natural level of generality, to be variations on a common theme. As just one example, consider the way an abstract formulation of the EM algorithm by Dempster et al. (1977) brought a sudden and clarifying unity to a vast array of applied problems and methods. The corresponding experience of revelation – literally a drawing back of the veil – that suddenly illuminates, after one has, at last and through effort, achieved an abstract understanding – can strike with all the sudden power of lightening. But just as the discharge in an electrical storm requires preparation through a gradual buildup of positive and negative poles, teaching for aesthetic effect takes time, because students cannot experience a surprise connection between A and B unless they have first come to understand each of A and B as separate and distinct. Hence the tension between efficiency and aesthetic.

For a caricature analog, imagine the choice between a gourmet meal at a fancy restaurant and a continuous IV drip of essential nutrients. The IV drip gets the necessary job done, with minimal claims on time and attention, but flavor and presentation are equally minimal.

A course about linear models offers many opportunities for surprise connections, although each such opportunity must be paid for with a nominal loss of efficiency. Example 5.1 offers a half-dozen instances. Each offers the possibility of a surprise connection between an (i) and a (ii). For each, the (i) is a standard element of the mainstream curriculum, and always taught. The (ii) is typically regarded as optional, sometimes taught, sometimes not. When it is taught, however, it is typically presented as an auxiliary consequence of (i). In order to teach the connection between (i) and (ii) as a surprise, it would be necessary to present (ii) independently and de novo, a choice that would ordinarily be declined as an unaffordable luxury.

Example 5.1 [Six possible surprise connections]

Least squares and projections. (i) Least squares estimators minimize the sum of squared residuals – the vertical distances from observed to fitted – and are found by setting derivatives to zero. (ii) The residual sum of squares is the squared distance from the response vector to the column space of the covariates; the least squares estimate is obtained by perpendicular projection.

Correlation and angle. (i) The correlation measures the goodness of (linear) fit. Invariance criteria dictate that the squared correlation is a suitably normalized residual sum of squares. (ii) The correlation is the cosine of the angle between the mean-adjusted response and covariate vectors.

Covariance and inner product. (i) When moments exist, the covariance is the integral of the product of the mean-adjusted response and covariate. (ii) If the usual moment assumptions hold (see Section 6), the covariance of two linear combinations is proportional to the usual Euclidean inner product of their coefficient vectors.

Least squares and Gauss-Markov estimation (i) Least squares estimators have minimum variance among linear unbiased estimators, with an algebraic proof. (ii) The linear unbiased estimators constitute a flat set; the estimator with minimum variance corresponds to the shortest coefficient vector in the flat set, which, like the least squares estimator, is obtained by orthogonal projection; see details in Section 8.

The multivariate normal density and the Herschel-Maxwell theorem (see details in Section 9) (i) The multivariate normal has density proportional to

$$\exp\left(-\frac{[\boldsymbol{Y}-\boldsymbol{\mu}]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{Y}-\boldsymbol{\mu}]}{2}\right).$$

The chi-square, t and F distributions are defined by their densities, and their relationships to the multinormal are derived by calculus. Calculus is also used to show that the multinormal is spherically symmetric, and that orthogonal components are independent. Standard results of sampling theory are derived by calculus. (ii) Given spherical symmetry and the orthogonality property which together define the normal, and given definitions of chi-square, t and F in terms of the multivariate unit normal, the sampling theory results can be derived without relying on densities.

The nested F-test and angle. (i) Given a full model $\boldsymbol{Y} = \beta_0\boldsymbol{x}_0+\beta_1\boldsymbol{x}_1+\boldsymbol{\varepsilon}$ and a reduced model $\boldsymbol{Y} = \beta_0\boldsymbol{x}_0 + \boldsymbol{\varepsilon}$, a test of the null hypothesis that the reduced model holds can be based on the F-ratio comparing the two residual sums of squares, as described in Section 4. (ii) Alternatively, a test can be based on the angle between the projections of the response vector onto the full and reduced model spaces.

With these examples as background, the argument for presenting the theory and practice of linear models in a way that celebrates their mathematical beauty is straightforward: The continued health and growth of our profession depends on attracting mathematically talented students who can rise to the methodological challenges and provide the unifying abstractions of the future. Many of these students we most need, especially the most mathematically talented, are attracted to mathematics for its own sake. If we present statistics as nothing more than applied data analysis, we may lose them to other subjects.

In this context, and in passing, it is worth noticing another tension: Applied data analysis, because it is anchored in context, tends to pull our profession apart, in opposing directions. If you choose to analyze data related to marketing, and I choose to analyze data from molecular biology, the more you and I devote ourselves to our separate areas of application, the less we have in common. In contrast to applied context, our profession's mathematical core is one of the things that hold us together. Even if you choose to do market research and I choose to work with microarrays, we both may well use generalized linear models or hierarchical Bayes. Prior to either of those, we both need a course in linear statistical models.

This section and the one before it might be seen as an argument for moving our teaching back toward mathematics, and so away from analyzing data, but that is not my intent. My conviction is that we can recognize the tension between data analysis and mathematics for its own sake without sacrificing one to the other. We can hope that our more mathematically talented students will aspire to be like the 19^{th} century pioneers Legendre and Gauss, who cared about solving scientific problems and cared about pure mathematics. Statisticians know Legendre and Gauss for their work with least squares; pure mathematicians know them for their work in number theory.

6. Two Organizing Principles for the Topics in a Course on Linear Models

Like Leo Tolstoy's happy families, almost all expositions of least squares follow the same general organization, according to the number of covariates in the model: Start with simple linear regression (one covariate), then move on to a treatment of models with two covariates, and from there to models with more than two covariates. (Some treatments skip the middle stage, and go directly from one covariate to two or more.) This organization-by-dimension echoes the way we traditionally order the teaching of calculus, first spending two semesters on functions of a single variable, and only then turning to functions of two or more variables.

Starting with simple linear regression (and one-variable calculus) offers the very major advantage that there are exactly as many variables in the model (one response plus one covariate) are there are dimensions to a blackboard or sheet of paper, which makes it comparatively easy to draw useful pictures. With two covariates, you need three dimensions, and pictures require perspective representations in the plane. With three or more covariates, you need to rely on training your mind's eye. A course organized by number of covariates fits well with the escalating difficulty of visualization; see Kleinbaum and Kupper (1978).

Despite the very real advantage based on visualization, I conjecture that the main reason for the near-universality of the usual organization-by-dimension comes from our prerequisite structure. We take it for granted that students in a least squares course have taken (and indeed, we assume, should have taken) at least one previous course in statistics. Such students will have seen simple linear regression already, so in accordance with an "overlap principle", it is sound pedagogy to start the least squares course with something that overlaps with what is already at least partially familiar. In short: If we assume students already know about simple linear regression, it makes sense to start their new course with simple linear regression.

Suppose, however, that your students have taken linear algebra, but have never taken probability or statistics before. Is the usual organization the best choice for these students? As context for thinking about this question, Figure 5 below shows a sense in which the content of a least squares course has a natural two-way structure.

Distribution assumptions	Number of covariates		
	One	Two	Three or more
A. None			
B. Moments			
C. Normality			

Figure 5. Two way structure of the content of a least squares course.

The rows of Figure 5 correspond to three different versions of the linear model:

(a) No distributional assumptions: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, with no assumptions about ε.
(b) Moment assumptions $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, with $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}$, that is, (a) $\mathrm{E}[\varepsilon_i] = 0$, (b) $\mathrm{Var}[\varepsilon_i] = \sigma^2$, (c) $\mathrm{Corr}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$.
(c) Normality assumption: $\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, that is, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, with $\mathrm{E}[\boldsymbol{\varepsilon}]= \mathbf{0}$,$\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}$, and each ε_i is normally distributed.

In the context of Figure 5, the standard organization for a least squares course is "one column at a time". (i) Start with simple linear regression, and move down the column. First, find the formula for the least squares slope and intercept. Then find the sampling distribution of the estimators, and use those for confidence intervals and tests. (ii) Having thus completed the left-most column, move to the middle column and repeat the process: estimators, sampling distributions, intervals and tests. (iii) At some point it is common to make a transition to matrix notation in order to allow a more compact treatment of the general case. This rough outline offers some flexibility about where to locate additional topics such as residual plots, transformations, influence, and multicollinearity, and different authors have different preferences, but the general reliance on this outline is near-universal.

An alternative organization for a linear models course is "one row at a time". (i) Start with no assumptions other than the linearity of the model and the fact that errors are additive. With no more than this it is possible to fit linear models to a whole range of data sets, with an emphasis on choosing models that offer a reasonable fit to data and context, as in Example 2.1. All four of the methodological challenges described in Example 2.2 can be addressed in this first part of a course. (ii) Next, add the moment assumptions. There are three main theoretical consequences: The moments of the least squares estimators, the expectation for the mean square error, and the Gauss-Markov theorem, discussed at length in Section 8. (iii) Finally, add the assumption that errors are Gaussian. At this point it becomes possible to obtain the usual sampling distributions, and to use those distributions for inference.

I see five important advantages to organizing a least squares course by strength of assumptions.

(a) Organizing by assumption follows history, taking what Toeplitz (1963) advocated as the "genetic" approach to curriculum. For least squares, the earliest work was distribution free. The moment and normality assumptions came later, after the least squares principle and solutions had taken root. As Toeplitz argues, often (though not automatically) what comes earlier in history is easier for students to learn.
(b) A course organized this way follows a "convergence principle", beginning from the least restrictive assumptions and most broadly applicable consequences, then narrowing in stages to the most restrictive assumptions and most narrowly applicable consequences.
(c) This organization gives students an immediate entrée to the challenge of model choosing. Good applied work in statistics almost always involves the creative process of choosing a good model, a process that is hard to teach within the narrow confines of simple linear regression, a context where the only y and x are both given. Put differently, starting with simple linear regression risks coming off as "spinach first, cake after" because the traditional ordering tends to emphasize the mechanical and technically difficult, postponing what is most interesting about analyzing data until much later in the course. Allowing multiple covariates from the start gives students an early taste of "the good stuff."

(d) Organizing by assumptions means that the first part of the course uses no probability, and involves no inference. For several weeks, the course can focus on the creative challenge of choosing models that offer a good fit to data and a meaningful connection to context, without the technical distractions of distributions, p-values, and coverage probabilities.

(e) Finally, and certainly not least, the organization reinforces the important logical connection between what you assume and what consequences follow. To the extent that we want to help our students learn to solve methodological problems, we owe it to them to make clear how what you assume determines what you can conclude.

As I see it, these advantages are relevant for any least squares course, but focus for the moment on the mathematics student who has taken matrix algebra but has not yet taken probability or statistics. For such a student, the algebraic aspect of working with several simultaneous variables is familiar ground. The more fortunate student may even be acquainted with the geometry of several dimensions. Probability and statistics, however, are new, and as experienced teachers know, probability is hard. For students new to the subject, being expected to learn all at once, at the start of a course, about continuous densities, probabilities as areas under curves, expected values and variances both as integrals and as measures of long-run behavior, not to mention the initially counter-intuitive post-dictive[5] use of probabilities for hypothesis testing – this is a lot to ask, even in the limited context of simple linear regression. Organizing a least squares course by assumptions introduces probability only after several weeks of working with linear models.

Moreover, probability is introduced in two stages, starting with moments only. By deferring sampling distributions and inference, the "moments-only" section of the course is able to focus on the difficult cognitive challenge of integrating the abstract mathematical description with an intuitive understanding of random behavior and the long run. Since moments can be understood as long-run averages, a "moments-only" section offers a gentler introduction to probability than does one that covers moments, sampling distributions, and inference all at once.

For the third part of the course, the basic results are pretty much standard: inference about β_j, inference about σ^2, and the F-test for comparing two nested models. Some courses may prefer to state and illustrate the results without deriving them, and among the many books that do present derivations, there is a variety of approaches. Some authors work directly with multivariate densities. Others (see, e.g., Hocking, 1996, Chapter 2) rely instead on moment generating functions. Toward the end of the next section, I describe yet a third approach, based on the Herschel-Maxwell theorem.

The next section provides more detail about a possible least squares course organized by assumptions.

7. One Model or Three?

The organization by assumptions described in the last section amounts to teaching linear models as a hierarchy of three classes of models based on assumptions about errors: none, moments only, and the normal distribution for errors. A common alternative is to work with a single class of models, those for which the conditional distribution of Y given X is normal with mean $X\beta$ and variance $\sigma^2 I$. Some books (see e.g. Ramsey and Schafer, 2002, p. 180) list all four assumptions of the model simultaneously. Draper and Smith (1966), on the other hand, is explicit that the assumptions need not come as a package. The first 16 pages of the book make no assumptions about errors. Then, on p. 17, the moment and

[5]Dempster (1971) distinguishes between the ordinary, predictive, use of probability and its retrospective, backward-looking, "postdictive" use to assess a model in the context of observed data.

normality assumptions are presented together for the one-predictor case. When the multivariable case is presented on p. 58, the moment and normality assumptions are included as part of a single package description of the model. The three kinds of estimation (minimizing the sum of squared residuals, minimizing the variance among linear unbiased estimators, maximizing the likelihood of the data) are clearly distinguished, but presented one after the other in the space of just two pages.

Other books (see e.g. Casella and Berger, 2002; Christensen, 1987; Searle, 1971; Terrell, 1999), whose emphasis is more mathematical, are clear and explicit that there are three models that correspond to three sets of increasingly strong assumptions and three corresponding sets of consequences, but, sadly in my opinion, books that emphasize data analysis tend not to be clear that there are three distinct sets of assumptions, while books that are explicit about the hierarchy of assumptions tend not to devote much time and space to modeling issues, in particular, to the connection between the applied context and the appropriate set of assumptions.

The question, "One model or three?" might be rephrased as a question about the origins of linear models, in the form "Least squares or regression?" As Stigler (1990) explains in detail, there are two different origins, separated by almost a century. The later of the two origins is cited by one of our best known expositions of linear models, (Neter et al., 1989, p. 26): "Regression was first developed by Sir Francis Galton in the latter part of the 19^{th} century". Some books, whether elementary (Freedman et al., 1978) or intermediate (Ramsey and Schafer, 2002), introduce regression models by means of conditional distributions of Y given X. Either, as with Galton, Y and X are assumed to have a joint bivariate normal distribution (see e.g. Freedman et al., 1978), or else the distribution of X is not specified but the conditional distribution of Y given X is Gaussian (see e.g., Ramsey and Schafer, 2002). For this approach, there is essentially one model, namely, that $\boldsymbol{Y}|\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

Least squares, in its earliest uses, began about a century earlier than Galton's work with the bivariate normal (Stigler, 1990), in a context where the methodological challenge was to find a "solution" to an inconsistent linear system of equations. In the context of the astronomical and geodesic challenges in the second half of the 1700s, there was no perceived need, and no recognized basis, for distributional assumptions. The least squares solution evolved, and then became established, long before Galton. Historically, then, we had linear models and least squares before, and independently of, any assumptions about the behavior of the errors of observation.

Because I have been unable to find a book on least squares whose organization follows the alternative path I have described, I will spell out in more detail the kind of organization I use in my own course on linear models. The course has three parts, of roughly six, three, and three weeks each (leaving a week for the inevitable slippage).

Part A [Linearity of the model, additivity of the errors: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$] For a course organized according to assumptions, the first part takes the deliberately naive view of data, that "what you see is all there is", i.e., there is no hidden "correct model" to be discovered, no unknown "true parameter values" to be estimated. The goal of the analysis is to find an equation that summarizes the pattern relating y values to x values, an equation that gives a good numerical and graphical fit to the data and good interpretive fit to the applied context, without being needlessly complicated.

The essential content for this part of a course has already been described in Section 2 in the context of the AAUP data of Example 2.1. There are four main clusters of topics, each driven by a methodological challenge: (i) Solving inconsistent linear systems; finding least squares estimates and the geometry of variable space; (ii) measuring goodness of fit and strength of linear relationships; simple, multiple, and partial correlation; adjusting for

a variable and partial residual plots; (iii) measuring influence and properties of the hat matrix; and (iv) measuring collinearity and the variance inflation factor.

As argued in Section 6, the pedagogical advantage of waiting to introduce probability is that you are thus able to focus in the first several weeks exclusively on issues of modeling, assessing fit, adjusting one set of variables for another set, and the influence of individual cases.

PART B [THE MOMENT ASSUMPTIONS: $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2$, $\mathrm{Corr}(\varepsilon_i, \varepsilon_j) = 0$ IF $i \neq j$] The three Moment Assumptions are a Trojan Horse of sorts: Hiding behind an innocent-seeming outer technical shell of probabilistic statements, these assumptions smuggle into our model the very strong implied assertion that, apart from random error and modulo the Box caveat,[6] the response we observe is in fact a true value whose exact functional relationship to the predictors is indeed known. To me, it is important to do all we can to impress upon our students how hard it should be to take these assumptions at face value.

Mathematically, the three main consequences of the moment assumptions are the moment theorem, that least squares estimators are unbiased with covariance matrix $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, the variance estimation theorem, that the expected residual mean square is σ^2, and the Gauss-Markov theorem, that among linear unbiased estimators, least squares estimators are best in the sense of having smallest variance. The first two theorems are direct, matrix-algebraic corollaries of the moment properties of linear combinations of vectors of independent random variables with mean 0, variance 1.

PROPOSITION Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$ satisfy $\mathrm{E}[Z_i] = 0$, $\mathrm{Var}[Z_i] = 1$, $\mathrm{Corr}(Z_i, Z_j) = 0$ if $i \neq j$, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ be vectors of constants and c a scalar constant. Then $\mathrm{E}[\boldsymbol{a}^\top \boldsymbol{Z} + c] = c$, $\mathrm{Var}[\boldsymbol{a}^\top \boldsymbol{Z} + c] = \|\boldsymbol{a}\|^2$, and $\mathrm{Cov}(\boldsymbol{a}^\top \boldsymbol{Z}, \boldsymbol{b}^\top \boldsymbol{Z}) = \boldsymbol{a} \cdot \boldsymbol{b}$.

COROLLARY [MOMENT THEOREM] For a linear model that satisfies the Moment Assumptions, $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

The proof is just a matter of applying the proposition to $\hat{\beta}_j = \boldsymbol{\mu}_j^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$, where μ_j is a vector of zeros except for a 1 in the jth position.

PROPOSITION Let $\boldsymbol{Z}$ be as above, and $\boldsymbol{A}$ an $n \times n$ matrix of constants. Then, $\mathrm{E}[\boldsymbol{Z}^\top \boldsymbol{A} \boldsymbol{Z}] = \sigma^2 \mathrm{tr}(\boldsymbol{A})$.

COROLLARY [VARIANCE ESTIMATION THEOREM] $\mathrm{E}[\boldsymbol{Y}^\top [\boldsymbol{I} - \boldsymbol{H}] \boldsymbol{Y}] = \sigma^2(n - p)$, where p is the rank of the hat matrix $\boldsymbol{H} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$.

The Gauss-Markov theorem, that least squares estimators are best among linear unbiased estimators, is typically given short shrift, if it gets any attention at all, but in my opinion this important result deserves much more attention than it ever gets. Accordingly, I have given it in this article a section unto itself, Section 8.

PART C [THE NORMALITY ASSUMPTION: THE ε_i ARE NORMALLY DISTRIBUTED] Whereas the moment assumptions specify a long-run relationship between observed values and an underlying (assumed) truth, and represent a major step up from the distribution-less model, adding the third layer, that distributions are Gaussian, is a comparatively minor escalation, for the usual two reasons: theory guarantees normality when samples are sufficiently large; and experience testifies that a suitable transformation can make many unimodal distributions close to normal. In short, if you've got the first two moments, normality can be just a matter of transformation and possibly a little more data.

[6] Essentially, all models are wrong; some are useful; see Box and Draper (1987, p. 424).

Although from a practical point of view, going from Part B to Part C is not a big step, I suggest that nevertheless, there are important pedagogical and curricular reasons to teach Parts B and C as very distinct and separate units.

The pedagogical reason for working first with just moments, and only later to tackle entire distributions, was addressed in Section 6: this organization allows a course to focus on how means and variances of linear combinations are tied both to long-run averages and to the geometry of variable space.

The main curricular reason to teach "moments first, distributions after" is, as all along, to highlight the connection between assumptions and consequences. Our third set of assumptions – that distributions are normal – allows us to assign probabilities to outcomes. From a practical standpoint, going from moment assumptions to normality is but one small step, but from a theoretical standpoint, it is a giant leap. If we can assign probabilities to outcomes, we can do three very important new things: we can choose estimates that maximize the postdictive probability of the data; we can use models to assign p-values to tail areas, and use these p-values to compare models; and we can associate a probabilistically-calibrated margin of error with each estimator. In short, the normality assumption opens the door to maximum likelihood estimation, hypothesis testing, and confidence intervals, none of which are possible if all we have are the first two moments.

This section, and the one before it, have presented some reasons to reconsider the way we organize our exposition of linear models. The next two sections raise questions that are independent of whether a course is organized by dimension or by assumptions: How much attention does the Gauss-Markov theorem deserve? How should we teach the sampling distribution theory we need for inference?

8. The Role of the Gauss-Markov Theorem

In compact acronymic form, the Gauss Markov theorem says that OLS = BLUE: the ordinary least squares estimator is best (minimum variance) within the class of linear unbiased estimators. In my opinion, the Gauss-Markov theorem offers a litmus test of sorts, a useful thought experiment for clarifying course goals and priorities. How much time does the result deserve in your treatment of least squares? Among books with an applied emphasis, most don't mention Gauss-Markov at all. Some books just state the result; a few state the theorem and give a quick algebraic proof, of the sort illustrated later in this section. No book that I am aware of, certainly no book suitable for a first course in statistics, devotes much time and space to the result, although, as I hope to persuade you, a geometric proof can offer students a deeper understanding of the remarkable connections among Euclidean geometry, statistics, and probability.

As I see it, a risk in presenting the Gauss-Markov theorem too quickly is that students will see the result as merely asserting a secondary property of least squares estimators, namely, that their variances have the nice feature of being as small as possible. What is at risk of getting lost is that there are two quite different sets of assumptions about data, each with its own approach to estimation. Assuming nothing more than linearity of the model with additive errors, we get least squares estimation as a method for solving inconsistent linear systems. Adding a set of very strong and restrictive assumptions about errors opens up an entirely different approach to estimation, starting from the infinite set of all unbiased linear estimators, then choosing the one(s) that minimize variance. On the surface, there is no reason to expect that the two approaches will always give the same estimators, and yet they do.

To me, the implication for teaching is clear. To motivate students to appreciate the importance of the Gauss-Markov theorem, we have to convince them, through concrete examples, that best linear unbiased estimation is in fact based on a logic very different from that of least squares. Using that logic to obtain Gauss-Markov estimators for specific examples is an effective way to emphasize the difference. Least squares is purely a method for resolving inconsistent linear systems, a method that makes no assumptions whatever about the behavior of errors apart from their additivity. Gauss-Markov estimation rests on very strong assumptions: For every observed value, there is an unobserved, unknown true value, and the differences between the observed and true values are random and uncorrelated, with mean zero, and constant SD.

These are such strong assumptions that it is not hard to persuade students that for such a very high price, we should get something better than mere least squares in return. So our class devotes half a week to finding best linear unbiased estimators, for a variety of simple, concrete examples, only to find that we always end up with estimators that are the same as the least squares estimators. Temporarily at least, the denouement should be a let-down: On the surface, our strong additional assumptions seem to buy us zero! However, on reflection, we can see the result as a surprise endorsement of least squares, in that the assumptions guarantee, via the moment theorem[7] that over the long run, least squares estimates average out to the true values, provided the model is correct.

Example 8.1 [Best linear unbiased estimators for the crystal problem] Consider again the problem of estimating the crystal width β from two observed values $Y_1 = \beta+\varepsilon_1$ and $Y_2 = 2\beta+\varepsilon_2$. (a) Linear unbiased estimators: Find the set of all $a{=}(a_1, a_2)^\top$ for which $\mathrm{E}[\boldsymbol{a}^\top \boldsymbol{Y}] = \boldsymbol{\beta}$. (b) "Best", which one(s) of the estimators in (a) give the smallest variance? Solution (see Figure 6):

(a) Linear unbiased estimators: Setting $\mathrm{E}[a_1Y_1 + a_2Y_2] = \beta$ leads to a single linear equation $a_1 + 2a_2 = 1$ whose solution is a flat set in $\mathbb{R}^2$. More generally, the coefficient vectors for the linear unbiased estimators are solutions to a non-homogeneous linear system, and so they form a flat set in $\mathbb{R}^n$.

(b) "Best": The moment assumptions imply that $\mathrm{Var}[a_1Y_1+a_2Y_2] = \sigma^2(a_1^2+a_2^2)$. More generally, given the usual moment assumptions, the variance of any linear estimator is proportional to the squared length of its coefficient vector. Thus the "best" linear unbiased estimator corresponds to the shortest of the vectors in (a).

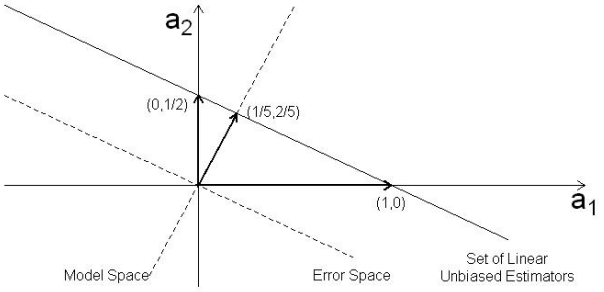

Figure 6. The linear unbiased estimators for the crystal problem Three linear unbiased estimators are shown as vectors: $(1, 0)$ is the estimator y_1; $(0, 1/2)$ is the estimator $y_2/2$; and $(1/5, 2/5)$ is the least squares estimator $(y_1 + 2y_2)/5$. The set of all linear unbiased estimators forms the solid line, which is parallel to error space. The shortest coefficient vector is the one with minimum variance.

[7] $\mathrm{E}[\hat{\beta}] = \beta,\ \mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$

After students have solved a few problems of this sort, they find it natural to ask about the general situation: Which subspace is it that gets translated in order to get the flat set of unbiased estimators? (Answer: It is always error space $\langle \boldsymbol{X} \rangle^\top$. Is there a quick way to find the shortest coefficient vector? (Answer: The shortest vector will always be perpendicular to error space.) Which is better: least squares or Gauss-Markov? (Answer: You'll be surprised.) In short, students are motivated, based on their experience, to understand the Gauss-Markov theorem, and to want to know why it is true.

To the extent that there is a "standard" proof of the Gauss-Markov theorem, it tends to be algebraic, essentially a variation on the derivation of the "computing rule" for the sample standard deviation: Complete the square, and show that the sum of cross-products is zero.

Here, to set the stage, is an abbreviated version of the elementary algebraic derivation of the "computing" formula for the sample variance:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} [(y_i - \bar{y}) + \bar{y}]^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + 2 \sum_{i=1}^{n} \bar{y}\,(y_i - \bar{y}) + \sum_{i=1}^{n} \bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + n\bar{y}^2.$$

In Neter et al. (1989, pp. 66-67), a similar proof is given that the least squares slope for simple linear regression is best among linear unbiased estimators: Write the least squares slope as $\sum k_i y_i$, and write an arbitrary linear unbiased estimator as $\sum c_i y_i = \sum (k_i + d_i) y_i$, where d_i is the difference between the least squares coefficient and that for the arbitrary estimator. Then, much as for the sample variance:

$$\text{Var}\left[\sum c_i Y_i\right] = \sigma^2 \sum c_i^2 = \sigma^2 \left(\sum k_i^2 + 2 \sum k_i d_i + \sum d_i^2\right) = \sigma^2 \left(\sum k_i^2 + \sum d_i^2\right).$$

The proof that the cross products sum to zero requires substituting for k_i from the formula for the least squares slope and doing some messy algebra, leading to $\text{Var}[\sum c_i Y_i] \geq \sigma^2 (\sum k_i^2)$; see Fox (1997, p. 127), Freedman (2005, p. 53), and Graybill (1961, pp. 115-116).

A matrix version of the same idea (see Terrell, 1999, p. 393) handles the general case. Let $\boldsymbol{AY}$ be an arbitrary linear unbiased estimator for β, and write $\boldsymbol{A} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top + \boldsymbol{D}$. Then,

$$\begin{aligned} \boldsymbol{A}\boldsymbol{A}^\top &= [(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top + \boldsymbol{D}][(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top + \boldsymbol{D}]^\top \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D}^\top + \boldsymbol{D}\boldsymbol{D}^\top \\ &\geq (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{D}^\top. \end{aligned}$$

Here, as before, the cross product terms vanish – algebraic proof required – and since $\boldsymbol{D}\boldsymbol{D}^\top$ is positive semi-definite, the result follows.[8]

I find these proofs instructive because of the way they echo a useful, recurring algebraic trick, but I do not find them illuminating in the sense of shedding bright light on deep ideas. Deep ideas in mathematics tend to come from abstraction-as-process, starting concretely, with simple examples, looking for patterns, gradually and systematically escalating the complexity of examples in order to see which patterns fall away and which others survive, and finally, discovering a general argument that explains why the patterns must be what they are.

Here are four ideas I consider deep, ideas that we can illuminate and reinforce via a proof of the Gauss-Markov theorem.

[8] An alternative proof using Lagrange multipliers is given in Hocking (1996, p. 97) and in Searle (1971, p. 88).

(a) Every least squares estimator is a linear function of the observed values, with coefficient vector in model space.

If the standard three moment assumptions hold, then:

(b) The SD of a linear combination is proportional to the length of its coefficient vector.
(c) Any linear estimator whose coefficient vector belongs to error space is an unbiased estimator of 0, and vice-versa: if the coefficient vector belongs to error space, the linear estimator is unbiased for 0.
(d) Every linear unbiased estimator is a Pythagorean sum of (a) the corresponding least squares estimator and (b) some unbiased estimator of 0.

For simplicity the proof that follows will consider only estimators of individual components β_j, but the results apply to any estimable $\boldsymbol{c}^\top\boldsymbol{\beta}$. Every linear estimator $\boldsymbol{a}^\top\boldsymbol{Y}$ can be identified with its coefficient vector $\boldsymbol{a}$. In what follows, $\boldsymbol{a}$ and $\boldsymbol{a}^\top\boldsymbol{Y}$ denote an arbitrary linear estimator; $\boldsymbol{a}_j$ and $\boldsymbol{a}_j^\top\boldsymbol{Y}$ denote an arbitrary linear unbiased estimator of β_j; $\hat{\boldsymbol{a}}_j$ and $\hat{\boldsymbol{a}}_j^\top\boldsymbol{Y}$ denotes the least squares estimator of β_j.

The proof of the Gauss-Markov theorem rests on the four facts (a)-(d) listed above. Each is truly important, each deserves individual attention, and each has its own one-line proof. After that, the Gauss-Markov result should be all but self-evident, with no algebraic trick required:

(a) The coefficient vectors for least squares estimators belong to model space: if $\hat{\beta}_j = \hat{\boldsymbol{a}}_j^\top\boldsymbol{Y}$, then $\hat{\boldsymbol{a}}_j \in \langle\boldsymbol{X}\rangle$. This is because $\hat{\beta}_j$ is the jth element of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$, so $\hat{\boldsymbol{a}}_j$ must be a linear combination of columns of $\boldsymbol{X}$.
(b) ("Error space lemma") The linear unbiased estimators of 0 correspond to error space: $\mathrm{E}[\boldsymbol{a}^\top\boldsymbol{Y}] = 0$ for all $\boldsymbol{\beta} \Leftrightarrow \boldsymbol{a} \in \langle\boldsymbol{X}\rangle^\top$. This is because $\mathrm{E}[\boldsymbol{a}^\top\boldsymbol{Y}] = \boldsymbol{a}^\top\boldsymbol{X}\boldsymbol{\beta}$, so $\boldsymbol{a}^\top\boldsymbol{X}$ must be 0.
(c) The linear unbiased estimators of β_j form a flat set parallel to error space: $\mathrm{E}[\boldsymbol{a}_j^\top] = \beta_j \Leftrightarrow \mathrm{E}[\boldsymbol{a}_j^\top\boldsymbol{Y} - \hat{\boldsymbol{a}}_j^\top\boldsymbol{Y}] = 0 \Leftrightarrow (\boldsymbol{a}_j - \hat{\boldsymbol{a}}_j) \in \langle\boldsymbol{X}\rangle^\perp$.

THEOREM [Gauss-Markov] Among all linear unbiased estimators for b_j, the least squares estimator has minimum variance.

PROOF Taken together, (a)-(c) establish that any linear unbiased estimator a_j is the hypotenuse of a right triangle whose perpendicular sides are the least squares estimator $\hat{\boldsymbol{a}}_j$ and the difference $(\boldsymbol{a}_j - \hat{\boldsymbol{a}}_j)$. Thus, $\|\boldsymbol{a}_j\|^2 \geq \|\hat{\boldsymbol{a}}_j\|^2$, as required. ■

9. Two Approaches to Sampling Distributions

Regardless of whether you choose to teach the moment assumptions and their consequences prior to, and separate from, the normality assumption and its consequences, or you decide instead to combine the two sets of assumptions and teach their consequences as part of the same integrated logical development, whenever you do eventually come to sampling distributions and their use for inference, you face an important choice: To what extent do you rely on probability densities and calculus-based derivations? To what extent might you want to avoid those?

The core content is the same either way. Students need to learn about t-tests and intervals for regression coefficients, about the scaled chi-square distribution for error mean square, and about the F-distribution for the ratio of mean squares used to test a nested pair of linear models. These goals set a multi-layered agenda of (i) four distributions to define,

(ii) five basic probabilistic relationships to establish, (iii) three key sampling distributions to derive (at a minimum), and (iv) the use of those sampling distributions for inference.

Four layers of the core agenda for Part C:

(a) Four distributions to define: multivariate normal, chi-square, t, and F.
(b) Five probability results:
  (b.1) Linear functions of multivariate normals are normal.
  (b.2) Projections of normals into orthogonal subspaces are independent.
  (b.3) The squared length of a standard normal vector (mean 0 and variance I) is chi-square.
  (b.4) A standard normal divided by the square root of an independent chi-square is t.
  (b.5) The ratio of two independent chi-squares over their respective degrees of freedom is F.
(c) Three essential sampling distributions
  (c.1) A studentized regression coefficient has a t distribution.
  (c.2) The residual sum of squares divided by the true variance is chi-square.
  (c.3) When the reduced model is true, the F-statistic for comparing full and reduced models has an F-distribution.
(d) Use of the sampling distributions for inference.

At the extremes, I see two competing approaches to the distribution theory. One approach, which for the convenience of having a label, I call "classical", defines the four distributions in terms of their probability densities. The five results of the second layer are then derived using classical, calculus-based, probability methods. For example, to show (c.1), write the product of densities for n unit normals, change to new variables including $u = z_1^2 + \cdots + z_n^2$, and integrate over the other variables to obtain the marginal density for u. This result, along with the other four, can then be applied to obtain the three sampling distributions of the third layer.

This "classical" approach sketched above relies heavily on change of variable methods, multiple integration, and often on generating functions as well. An alternative, which I call here the "Herschel-Maxwell" approach after the theorem that provides logical justification, makes it possible to avoid calculus entirely. Astronomer Herschel (1850) and physicist Maxwell (1860) are credited with the two-dimensional and three-dimensional versions of the theorem that bears their names:[9]

THEOREM [HERSCHEL-MAXWELL] Let $\boldsymbol{Z} \in \mathbb{R}^n$ be a random vector for which (i) projections into orthogonal subspaces are independent and (ii) the distribution of $\boldsymbol{Z}$ depends only on the length $\|\boldsymbol{Z}\|$. Then $\boldsymbol{Z}$ is normally distributed.

If your students are already familiar with multivariate probability densities from a previous probability course, then establishing the converse of the theorem, proving (i) and (ii) above starting from the multivariate normal density, turns out to be easy. However, if you are teaching this material – inference for least squares – to students who have not yet seen probability densities, and who may not have seen any multivariable calculus either, the "classical" approach will be a stretch, at best. You can always finesse the distribution theory as "beyond the scope of this course", and many good books with an applied emphasis choose to do just that. But when I am teaching a course for students of mathematics, although I may be willing to omit an occasional proof, I would be embarrassed to present key sampling distributions as a set of black boxes.

[9]See Jaynes (2003, pp. 201-201) for Herschel's argument for the two-dimensional case, which Maxwell later extended to three dimensions.

Fortunately, the Herschel-Maxwell theorem justifies using properties (i) and (ii) as a definition, which can then lead to a simple, straightforward, density-free exposition. To start, let a random vector $\boldsymbol{Z} \in \mathbb{R}^n$ have a distribution that is both spherically symmetric (any rigid rotation leaves its distribution unchanged) and "ortho-independent" (projections into orthogonal subspaces are independent), and suppose the components Z_i have variances equal to 1. Thanks to Herschel-Maxwell, these properties can serve to define the standard n-dimensional multivariate normal. The general normal family then comes from taking linear transformations of $\boldsymbol{Z}$, and, with appropriate care about uniqueness, properties (c.3)-(c.5) above can be taken as definitions of the chi-square, t, and F distributions: A chi-square distribution is the distribution of $\|P_U(\boldsymbol{Z})\|^2$, where P_U is the orthogonal projection onto a subspace U whose dimension equals the degrees of freedom. The statistics $\boldsymbol{u}^\top \boldsymbol{Z}/\sqrt{\|P_U(\boldsymbol{Z})\|^2/\dim(U)}$ and $\big[\|P_U(\boldsymbol{Z})\|^2/\dim(U)\big]/\big[\|P_V(\boldsymbol{Z})\|^2/\dim(V)\big]$ follow t and F distributions, respectively, where $\boldsymbol{u}$ is a unit vector orthogonal to the subspace U and U and V are orthogonal subspaces.[10]

The necessary sampling distributions follow quickly from the definitions. (i) Since $\hat{\boldsymbol{\beta}}$ is a linear function of normals, its distribution is normal. (ii) Since $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\varepsilon}}$ lie in orthogonal subspaces, they are independent, and since $\hat{\boldsymbol{\beta}}$ is a function of $\hat{\boldsymbol{Y}}$ and the mean squared error is a function of $\hat{\boldsymbol{\varepsilon}}$, it follows that $\hat{\boldsymbol{\beta}}$ and MSE are independent. (iii) Apart from scaling, MSE is the squared length of a projection of a normal vector into error space, from which the chi-square distribution for SSE/σ^2 follows; similarly for the t distribution of a studentized regression coefficient. (iii) For testing whether the reduced model in a pair of nested linear models is "true", variable space is decomposed into three mutually orthogonal subspaces: error space, reduced model space (column space for the smaller model), and "difference space" (the orthogonal complement of reduced model space in full model space). Projections into these subspaces are independent. Under the null hypothesis that the reduced model is true, both the numerator and denominator in the usual F-test have scaled chi-square distributions, and so the ratio of their mean squares has an F-distribution.

Which approach is better, "Classical" or "Herschel-Maxwell?" Of course instructor tastes and priorities differ, course goals differ, and student backgrounds differ, but for me, tying the sampling distributions directly to the geometry of variable space highlights the connections that matter most for a deep understanding of the ideas. This is not at all to say that the calculus is a bad thing to teach, only that we should not allow calculus to usurp exclusive right-of-way to the sampling distributions we need for working with linear models, and so calculus should not be allowed to dictate how and when we teach linear models. Even if you prefer the calculus-based approach, the existence of the alternative I have just sketched demonstrates that a least squares course need not wait until after a multivariable-calculus-based probability course. Granted, for a first course in statistics, we may not want to take time in the statistics course to teach the needed calculus to students who haven't seen it before, and traditionally, teaching linear models to mathematical students has too often seemed to require all that calculus. But since we can teach least squares without the calculus, we are at the least logically free to reconsider how soon we might want to teach a course in linear models. In particular, teaching linear models as a first course remains an option.

[10]This way of defining the descendant distributions of the normal is in no way original with me, and has long been part of "the folklore". Published variants include Rice (1995), which defines the descendant distributions in terms of the normal and then derives their densities using calculus-based methods.

10. Linear Models: A First Statistics Course for Mathematics Majors?

In this final section I compare what I see as the advantages and disadvantages of the linear models course with those of the two standard introductions to statistics: (i) the lower level course with a an applied orientation, perhaps followed by a second applied course with an emphasis on modeling, and (ii) some variant of the standard two-semester upper division sequence in probability and mathematical statistics.

The applied introduction has the two main advantages of requiring little preparation in mathematics, and, in the better of its current incarnations, doing a good job of introducing students to the basics of data analysis. However, this course is not (and has never claimed to be) an introduction designed for students of mathematics. It has little mathematical substance, no explicit methodological challenges, little to appeal to a mathematical aesthetic, and no basis for carrying credit as an upper-level elective in a mathematics major. None of this is a criticism of the course itself; it was never intended to do the things I have just complained about.

I note in this context that the recent book Investigating Statistical Concepts and Methods (Rossman and Chance, 2006) does offer an applied introduction explicitly designed for students of mathematics. This book is brilliantly conceived and effectively implemented based on research about how students learn statistics, with interesting data sets and a thoughtful reworking of the usual introductory content in a way that makes it a much better match for mathematically motivated readers. Even this book, however, despite its originality, is in my judgment limited by its choice of content. If you start from the premise that you want to teach a variant of the usual introductory curriculum to mathematically strong students, it is hard to imagine a better approach than the one Rossman and Chance have given us. But for me the question remains: If you are designing a course for mathematics majors, why start with content from a course designed for a different readership? Why not start from scratch, with content chosen specifically to introduce mathematical students to statistics?

The standard sequence in probability and mathematical statistic has the one salient advantage of offering good practice with multivariable calculus, but I find little to recommend the typical course as a first statistics course for any student. Whereas the introductory applied course has been evolving steadily and in healthy ways since the 1960s, the main stream upper division sequence in probability and mathematical statistics has changed very little. The probability course can be, and often is, a course of great value, but probability is a branch of mathematics in a way that statistics is not, and so, from the point of view of introducing statistics, the probability course is yet another mathematics prerequisite that delays the first statistics course. The mathematical statistics course, like the probability course, offers practice with multivariable calculus, but all too often offers little else beyond its venerable parade of geriatric definitions and theorems, feebly searching for points of contact with modern practice.

Regarded as an introduction to statistics, the usual sequence has numerous shortcomings. Surely one of the most restrictive is the combination of prerequisite structure and the fact that in the typical year-long sequence, statistics comes so late. Although it is possible to teach mathematical statistics while requiring only three semesters of previous work in mathematics (two semesters of calculus plus probability), the typical course does a lot of multiple integrals, and the typical student would be better off starting with an additional semester of calculus.

Besides, how many year-long introductions to "X" do you know that defer talking about "X" until after an entire semester of getting ready? Would economists or physicists teach a year-long introduction to their subject with a first semester devoted exclusively to mathematics?

Another shortcoming is that the traditional mathematical statistics course remains devoted to problems with closed-form analytic solutions, e.g., finding the maximum likelihood estimator of a Poisson mean from a simple random sample, at a time when statistical practice relies increasingly on computer-intensive approximate methods.

Still another major shortcoming is that mathematical statistics, by its nature, cannot be about the analysis of data in any deep sense. We know, thanks to the inspiring examples of Larsen and Marx (1986), and Rice (1995), how to incorporate real data and authentic examples into our teaching of mathematical statistics. But, important as these real examples are, they are essentially illustrations of theory and methods, not invitations to detective work with data. Each data set is chosen to illustrate a particular given model; the model itself is given. Contrast this situation, with model given and data set chosen accordingly, to Example 2.1, where the data set is given and a variety of models will be considered in order to obtain a good fit and answer questions of interest. Of course future statisticians need and should learn the content of mathematical statistics, but for a first course, before students have had a chance to decide whether they might in fact be future statisticians, we should offer them an introduction that does a better job of showing the kinds of things that statisticians do in practice.

Finally, one should ask of our introductory statistics courses, "If this is a student's first statistics course, where does it lead?" The applied first course leads naturally to an applied second course, ideally one with a modeling emphasis, as in Ramsey and Schafer (2002) or Cannon et al. (2010). The probability course leads naturally to mathematical statistics as a first statistics course, but what about a second statistics course? Although various modeling courses are possible as second courses, these do not make essential use of much of the content of the mathematical statistics course, and could have been offered instead of it, as an immediate follow-up to the probability course. The only statistics course that makes essential use of the core content from the standard first course in mathematical statistics is a second, graduate-level course in mathematical statistics, and when it comes to data analysis, that course suffers from the same shortcomings as the first course.

It is important for me to be clear. When I direct my complaints against the *usual* mathematical statistics course as a *first* course in statistics, the two italicized words deserve emphasis. (If I appear to be complaining that pizza doesn't taste like lobster, I have failed in my exposition.) "*First*": For many students, some version of the mathematical statistics course might be a very useful and engaging follow-up to some other first course that does justice to applied data analysis. "*Usual*": Fortunate students may have a teacher who chooses to teach an unusual version of the mathematical statistics course. I don't want to take the time and space in this article to review the options in detail, although there are now enough original books that I think our profession would benefit immensely from a comparative review of that sort. For brevity, I will limit myself here to two main points, followed by a quick summary of some of the most original books.

Two important facts about textbooks for mathematical statistics:

- The convex hull of content, organization, and approach has been expanding with Hubble-like acceleration. I take this rapid curricular expansion as a healthy symptom of our profession's growing discontent with the usual course.
- Despite our rapidly expanding horizon of options, time remains an inelastic, zero-sum quantity, and so it is in the nature of the content of probability and mathematical statistics that any focus on abstract theory of inference takes time and attention away from the challenge of modeling data relationships.

Caveats duly deployed, here are capsule salutes to some particularly impressive innovations, in chronological order:

- Larsen and Marx (1986). Although conventional in content by today's standards, this was the first mathematical statistics book to make effective use of real data carefully chosen for intrinsic interest.
- Rice (1995). Much more mathematical, more demanding, and more statistically sophisticated than Larsen and Marx (1986), this book is also a pioneer in its use of real data.
- Terrell (1999). This book broke new ground by plowing up the usual order of topics in order to start with chapters on "Structural Models for Data" (Chapter 1: one and two-way layouts, simple and multiple regression, contingency tables, and logistic regression), and least squares methods (Chapter 2), before five chapters on probability through sampling distributions. This book is written at a comparatively high level, and does comparatively less with modeling than the topics of its first two chapters might seem to promise, but ten years ago it was far ahead of its time in putting two chapters on statistical topics ahead of probability.
- Nolan and Speed (2000). This is the most radical departure from the mainstream of any in my list. The authors use a sequence of applied case studies as a vehicle for teaching mathematical statistics. If you teach mathematical statistics and don't yet know this book, I urge you to challenge yourself by reading it, and thinking about how to use it in your teaching; see also Nolan (2003).
- Rossman and Chance (2006). Investigating statistical concepts and methods. This book is unique in the innovative way it rethinks and reshapes the usual introductory applied course to suit mathematically inclined students, using many interesting real data sets along the way.
- Lavine (2009); see also first edition (2006). You can get this book free from the web. It would be unfair to say it is worth every penny. In fact, it is worth the time and attention of anyone who teaches mathematical statistics to read this remarkable book carefully and to think hard about the way it challenges us to reconsider what we teach. If you have a year for your course, you have students who are skilled users of multivariable calculus, and you want to cover the content of the usual two-semester sequence but you also want to "focus on ideas that statisticians care about", this book does all that and much more. It is hard to condense my enthusiasm into just a few sentences that might persuade others, but consider the content: Probability is covered as needed, in chapters 1, 4 and 5 (of 8), to support an approach that puts likelihood front and center. Chapter 2 is an overview of "Modes of Inference" – direct use of likelihood, estimators and sampling distributions, Bayesian methods, prediction, and testing. Chapter 3 follows with a treatment of normal regression models. Traditional mathematical statistics waits until the final Chapter 8. Meanwhile, we get a chapter on Bayesian methods (including MCMC, both Metropolis and Gibbs), and another chapter on additional models (random effects, time series, and survival analysis). The mathematical demands are high, but the return is comparably high. A student who master's Lavine's book has an excellent foundation for advanced work in statistics; see MacEachern (2006).
- Chihara and Hesterberg (2011). This book assumes a semester of probability, but is much more applied than the usual book on mathematical statistics. After a quick probability review (Chapter 1), it begins with graphical approaches to data exploration (qq-plots and scatter plots, Chapter 2), introduces testing via randomization (Chapter 3), followed by chapters on sampling distributions and the bootstrap, leading to chapters on estimation, confidence intervals, least squares, and Bayesian inference. It is less demanding mathematically, more applied in its emphasis, and more modern in content than the usual book, which makes it a good choice if you want a modern applied book at the level of Larsen and Marx (1986).
- Pruim (2011). Like Chihara and Hesterberg (2011), this book combines material from probability and mathematical statistics with serious attention to applications. The open-

ing example is Fisher's "lady tasting tea", Chapter 1 introduces R and data summaries, both graphical and numerical, Chapter 2, on probability, has a section on hypothesis tests and p-values, and Chapter 3, on continuous distributions, has sections on kernel density estimators and quantile-quantile plots. The first half of the book ends with a chapter on estimation and testing, which includes one-sample methods for proportions and means. The second half of the book starts with a chapter on likelihood-based methods, then concludes with two chapters on linear models. In its emphasis on linear models, the book is similar to Terrell (1999), but Pruim's exposition is less demanding mathematically, and differs also in the use of R throughout, and the placement of the linear models at the end of the book. A course based on this book would give students a much better sense of data analysis and modeling than they would get from a more traditional course.

- Horton (2010). Horton has not yet written a textbook for mathematical statistics, but his description of a course using the "modified Moore method" and R is a pioneering example we can all learn from.

Despite my deep admiration for all of these innovative efforts, many are by default not suitable for the course I am advocating: a one-semester introduction to statistics that does not require multivariable calculus or probability, a course designed to attract sophomore mathematics majors to statistics, a course with substantial mathematics for its own sake, a course that mixes methodological challenges with data modeling. Mathematical statistics, by the nature of its content, is not suited to meeting these goals.

A linear models course is different. Not only can it be taught independently of a probability course, and without relying on change of variables or multiple integrals, but, in addition, the centrality of linear models within statistics, like the centrality of a wheel hub, offers radial paths in many directions. After taking a course in linear models, a student is ready for a course on correlated data, or time-to-event modeling, or generalized linear models, or time series, or Bayesian data analysis. Alternatively, the work with the geometry of n-space can lead to data analysis in function spaces.

After more than two centuries, least squares and linear models remain at the core of statistics, central to our practice, and central to our theory. Shouldn't we make linear models equally central to our curriculum?

Appendix 1: Possible Course Outline (see Section 7 for details)

(a) No assumptions. What you see is all there is:
  (a.1) Combining observations: "solving" (reconciling) inconsistent linear systems least squares theorem.
  (a.2) Measuring fit and strength of relationship.
  (a.3) Measuring influence.
  (a.4) Measuring collinearity.

(b) Moment assumptions. Errors are uncorrelated, with mean 0 and constant SD:
  (b.1) Moment theorem.
  (b.2) Variance estimation theorem.
  (b.3) Best linear unbiased estimation.
  (b.4) Gauss Markov theorem.

(c) Normality assumption. Errors are normals:
  (c.1) Herschel-Maxwell theorem and the normal distribution.
  (c.2) Distribution of OLS estimators.
  (c.3) t distribution and confidence intervals for β_j.
  (c.4) Chi-square distribution and confidence intervals for σ^2.
  (c.5) F distribution and the general regression significance test (nested F test).

Appendix 2: AAUP Data (Bellas and Reskin, 1994)

Subject	AvAcSal	PctFem	PctUnemp	PctNonAc	MedNASal
Dentistry	44214	15.7	0.1	99.4	40005
Medicine	43160	25.5	0.2	96.0	50005
Law	40670	34.0	0.5	99.3	30518
Agriculture	36879	12.9	0.8	43.4	31063
Engineering	35694	4.6	0.5	65.5	35133
Geology	33206	13.5	0.3	58.1	33602
Chemistry	33069	16.2	1.1	61.9	32489
Physics	32925	7.2	1.2	40.7	33434
LifeSciences	32605	29.8	1.4	27.4	30500
Economics	32179	14.8	0.3	34.2	37052
Philosophy	31430	23.1	1.8	17.1	18500
History	31276	30.5	1.5	20.5	21113
Business	30753	27.1	1.9	98.9	20244
Architecture	30337	31.6	1.1	98.3	21758
Psychology	29894	45.5	1.1	51.0	30807
EducPsych	29675	49.5	0.9	82.5	20195
SocialWork	29208	80.0	1.9	98.6	16965
Mathematics	29128	15.4	0.4	17.9	32537
Education	28952	48.1	0.7	97.1	19465
SocAnthro	27633	40.9	2.1	12.8	21600
Art	27198	57.6	2.1	96.6	11586
Music	26548	45.5	4.3	98.5	16193
Journalism	25950	52.3	3.1	93.4	20135
English	25892	53.6	1.9	12.1	18000
ForeignLang	25566	55.0	3.5	12.1	20352
Nursing	24924	94.2	1.4	96.2	17505
Drama	24865	58.5	5.5	97.1	20005
LibraryScience	23658	82.2	1.1	78.9	15980

Where AvAcSal = mean academic salary; PctFem = percentage of faculty who are women; PctUn = percentage unemployed; PctNonAc = percentage of jobs that are not academic; MedNASal = median non-academic salary.

References

Bartlett, M.S., 1933. The vector representation of a sample. Proceedings of the Cambridge Philosophical Society, 30, 327-340.

Bellas, M., Reskin, B.F., 1994. On comparable worth. Academe, 80, 83-85.

Box, G.E.P., Draper, N.R., 1987. Empirical Model-Building and Response Surfaces. John Wiley & Sons, New York.

Box, G.E.P., Hunter, W.G., Hunter, J.S., 1978. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley & Sons, New York.

Breiman, L., 1969. Probability and Stochastic Processes with a View Toward Applications. Houghton-Mifflin, Boston.

Bryant, P., 1984. Geometry, statistics, probability: variations on a common theme. The American Statistician, 38, 38-48.

Cannon, A.R., Cobb, G.W., Hartlaub, B.A., Legler, J.M., Lock, R.H., Moore, T.L., Rossman, A.J., Witmer, J.A., 2010. Stat 2: A Second Course in Undergraduate Statistics. Freeman, New York (accepted for publication).

Casella, G., Berger, R.L., 2002. Statistical Inference. 2nd edition. Duxbury, Pacific Grove, California.

Chihara, L., Hesterberg, T., 2011. Mathematical Statistics with Resampling and R. John Wiley & Sons, New York (to appear).

Christensen, R., 1987. Plane Answers to Complex Questions. Springer-Verlag, New York.
Dempster, A.P., 1968. Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading, Massachusetts.
Dempster, A.P., 1971. Model searching and estimation in the logic of inference. In Godambe, V.P., Sprott, D.A., (eds.). Foundations of Statistical Inference, Holt, Rinehart and Winston, Toronto, pp. 56-78.
Dempster, A.P., Laird, N.M., Rubin D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of The Royal Statistical Society Series B - Statistical Methodology, 39, 1-38.
Draper, N.R., Smith, H., 1966. Applied Regression Analysis. John Wiley & Sons, New York.
Eaton, M.L., 2007. William H. Kruskal and the development of Coordinate-Free Methods. Statistical Science, 22, 264-265.
Fox, J., 1997. Applied Regression Analysis, Linear Models, and Related Methods. Sage, Thousand Oaks, California.
Fraser, D.A.S., 1958. Statistics: An Introduction. John Wiley & Sons, New York.
Freedman, D.A., 2005. Statistical Models: Theory and Practice. Cambridge University Press, Cambridge.
Freedman, D.A., Pisani, R., Purves, R., 1978. Statistics. Norton, New York.
Graybill, F.A., 1961. An Introduction to Linear Statistical Models – Volume 1. McGraw-Hill, New York.
Herschel, J.F.W., 1850. Quetelet on probabilities. Edinburgh and Quarterly Reviews, 92, 1-57.
Herr, D.G., 1980. On the history of the use of geometry in the general linear model. The American Statistician, 34, 43-47.
Hoaglin, D.C., Welsch, R.E., 1978. The hat matrix in regression and ANOVA. The American Statistician, 32, 17-22.
Hocking, R.R., 1996. Methods and Applications of Linear Models. John Wiley & Sons, New York.
Hoel, P., 1947. Introduction to Mathematical Statistics. John Wiley & Sons, New York.
Hogg, R.V., Craig, A.T., 1959. Introduction to Mathematical Statistics. MacMillan, New York.
Horton, N.J., 2010. I Hear, I Forget. I Do, I Understand: Using R as the Centerpiece of a Modified Moore-Method Mathematical Statistics Course. In progress.
Horton, N.J., Brown, E.R., Qian, L., 2004. Use of R as a toolbox for mathematical statistics exploration. The American Statistician, 58, 343-357.
Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge.
Johnson, R.W., 1994. Estimating the size of a population. Teaching Statistics, 16, 50.
Kaplan, D.T., 2009. Statistical Modeling: A Fresh Approach. CreateSpace, Seattle.
Kleinbaum, D.G., Kupper, L.L., 1978. Applied Regression Analysis and Other Multivariate Methods. Duxbury, North Scituate, Massachusetts.
Kruskal, W.G., 1961. The coordinate-free approach to Gauss–Markov estimation and its application to missing and extra observations. Proc. Fourth Berkeley Symposium on Mathematical Statististics and Probability. University California Press, Berkeley, pp. 435-451.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Kuhn, T.S., 1977. The Essential Tension: Selected Studies in Scientific Tradition and Change. University of Chicago Press, Chicago.

Larsen, R.J., Marx, M.L., 1986. An Introduction to Mathematical Statistics and Its Applications. 2nd edition, Prentice-Hall, Englewood Cliffs, New Jersey.
Lavine, M., 2009. Introduction to Statistical Thought. Orange Grove Texts Plus, FL, USA. Available at http://www.math.umass.edu/~lavine/Book/book.html.
MacEachern, S.N., 2006. Review of Introduction to Statistical Thought. A publication from Journal of the American Statistical Association, USA.
Maxwell, J.C., 1860. Illustration of the dynamical theory of gases. Part 1. On the motions and collisions of perfectly elastic spheres. Philosophical Magazine, 19-32.
Moore, T.L., 1992. Getting more data into theoretical statistics courses. PRIMUS, 2, 348-356.
Neter, J., Wasserman, W., Kutner, M.H., 1989. Applied Linear Regression Models. 2nd edition, Irwin, Homewood, Illinois.
Nolan, D.A., 2003. Case studies in the mathematical statistics course. Science and Statistics: A Festschrift for Terry Speed. IMS Press, Fountain Hills, Arizona, pp. 165-176.
Nolan, D., Speed, T., 2000. Stat Labs: Mathematical Statistics through Applications. Springer-Verlag-Verlag, New York.
Olkin, I., Gleser, L.J., Derman, C., 1980. Probability Models and Applications. Macmillan, New York.
Pruim, R., 2011. Foundations and Applications of Statistics. AMS, Providence (to appear).
Ramsey, F.L. Schafer, D.W., 2002. The Statistical Sleuth: A Course in Methods of Data Analysis. 2nd edition, Duxbury, Pacific Grove, California.
Rao, C.R., 1965. Linear Statistical Inference and Its Applications. John Wiley & Sons, New York.
Rice, J.A., 1995. Mathematical Statistics and Data Analysis. 2nd edition, Duxbury, Belmont, California.
Rossman, A., Chance, B., 2006. Investigating Statistical Concepts and Methods. Duxbury, Pacific Grove, California.
Saville, D.J., Wood, G.R., 1991. Statistical Methods: The Geometric Approach. Springer-Verlag-Verlag, New York.
Scheffe, H., 1959. The Analysis of Variance. John Wiley & Sons, New York.
Searle, S.R., 1971. Linear Models. John Wiley & Sons, New York.
Stigler, S.M., 1990. The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press, Cambridge, Massachusetts.
Terrell, G.C., 1999. Mathematical Statistics: A Unified Introduction. Springer-Verlag, New York.
Toeplitz, O., 1963. The Calculus: A Genetic Approach. University of Chicago Press, Chicago.