

ASSESSING INEQUITY IN GRADING *

BY ROBERT J. VANDERBEI, GORDON SCHARF AND DANIEL MARLOW

Princeton University[†]

In this paper we describe a statistical procedure to account for differences in grading practices from one course to another. The goal is to define a course “inflatedness” and a student “aptitude” that best captures ones intuitive notions of these concepts.

1. Introduction. This paper is about statistical techniques for inferring which courses are more grade-inflated relative to others based on information gleaned *a posteriori* from actual student grades. Suppose a student takes both course X and course Y and gets a higher grade in course X than in course Y. Based on just one student, it is likely that the student simply has more aptitude for the material in course X than for the material in course Y. But, if most students who took both courses X and Y got a better grade in course X than in course Y, then one begins to think that course X simply employed a more inflated grading scheme.

Consider for example, a school with only four students, John, Paul, George, and Ringo. Suppose further that this school only offers six different courses from which the students select four to take. The students made their selections, took the courses, and we now have grading information as shown in Table 1. From this table, we see that George and Paul have received the same grades (in different courses) and so their *grade-point averages*, i.e. their *GPA*’s, are the same. Furthermore, John’s grades are only slightly better and Ringo grades only slightly worse than average. But, it is also clear that the Math class gave lower grades than the Economics course. In fact, there is a linear progression in grade-inflation as one progresses from left to right across the table. Taking this into account, it would seem that John took “harder” courses than Paul (the quotes are to emphasize that a course that gives lower grades is not necessarily more difficult even though we shall use such language throughout this paper), who took harder courses than George, who took harder courses than Ringo. Hence, GPA does not tell the whole story. John did the best in all of his courses, in many cases by a wide margin. Ringo, on the other hand, did the worst in all of his classes, again

*

AMS 2000 subject classifications: Primary 62J05; secondary 62P25

Keywords and phrases: mean, median, least squares, least absolute deviations

	MAT	CHE	ANT	REL	POL	ECO
John	B-	B	B+	A-		
Paul	C+	B-		B+	A-	
George		C+	B-		B+	A-
Ringo			C+	B-	B	B+

TABLE 1

Grading data from Beattie University. The six courses are Math (MAT), Chemical Engineering (CHE), Anthropology (ANT), Religion (REL), Politics (POL), and Economics (ECO).

by a wide margin. It is clear that John is a much better student than Ringo. Better to a degree that is not reflected in their GPA's.

Our aim is to develop a model that can be used to infer automatically the sort of conclusions that we have just drawn for this small example. Of course, one must consider the simplest suggestion of just computing averages within each course. Clearly, in Table 1, the Math course gave grades a full letter grade lower than the Econ course. One could argue that that is all one needs—just correct using average grades within each course. But, one can easily modify the simple example shown in Table 1 to make all the courses have the same average grade and all of the students have the same GPA but for which there is an obvious trend in the true aptitude of the students. Table 2 shows one rather contrived way to do this (using an unbounded list of courses and students).

Finally, the model must be computationally tractable so that it can be run for a school with thousands of students taking dozens of courses (over four years) selected from a catalogue of hundreds of courses.

2. The Model. We assume that there are m students and n courses. Our data consists of the grades for all courses taught. For each course, we assume that we have grading data for every student who took that course. But, we do not assume that every student takes every course offered. In fact, we assume quite the opposite, namely, that each student only takes a small sample of the complete suite of course offerings.

We assume that each student has an aptitude μ_i , $i = 1, 2, \dots, m$, which is unknown to us and which we wish to estimate, and that each course has an inflatedness ν_j , $j = 1, 2, \dots, n$, which is also unknown to us and also of interest to estimate. We assume that each grade X_{ij} can be approximated as the sum of the student's aptitude plus the course's inflatedness:

$$(2.1) \quad X_{ij} = \mu_i + \nu_j + \epsilon_{ij}, \quad (i, j) \in \mathcal{G}$$

	...	MAT	CHE	ANT	REL	POL	ECO	HIS	...
⋮	⋮	⋮	⋮	⋮					
John		B-	B	B+	A-				
Paul			B-	B	B+	A-			
George				B-	B	B+	A-		
Ringo					B-	B	B+	A-	
⋮						⋮	⋮	⋮	⋮

TABLE 2

A school with an infinite number of students and an infinite selection of courses. Every student has the same GPA and every course has the same course average. Yet, John is smarter than Paul is smarter than George is smarter than Ringo and Math is harder than Chemical Engineering is harder than Anthropology etc.

where \mathcal{G} represents the set of student-course pairs (i, j) for which we have a grade (i.e., student i actually took course j). And, of course, the $\epsilon_{i,j}$'s are the "errors" one needs to add to make the approximation an equality.

Ideally, grades should reflect aptitude. Hence, we would like to say that a student with a B-level aptitude should be expected to get B-level grades. In other words, inflatedness should measure deviations, both positive (for courses with high grades) and negative (for courses with low grades), around some neutral average grade. In other words, we wish to impose the added constraint that

$$(2.2) \quad \sum_j \nu_j = 0.$$

This, of course, is by choice. We need some sort of normalization. Without one, we could add an arbitrary constant to every μ_i and subtract the same constant from every ν_j without changing any of the $\epsilon_{i,j}$'s.

Our aim is to find the best "fit" to the data. That is, we wish to choose the μ_i 's and the ν_j 's in such a manner as to make the $\epsilon_{i,j}$'s as small as possible. To do this, we minimize the sum of the squares of the $\epsilon_{i,j}$'s:

$$(2.3) \quad \begin{aligned} &\text{minimize} && \sum_{(i,j) \in \mathcal{G}} \epsilon_{i,j}^2 \\ &\text{subject to} && X_{ij} = \mu_i + \nu_j + \epsilon_{ij} \quad \text{for } (i, j) \in \mathcal{G} \\ &&& \sum_j \nu_j = 0. \end{aligned}$$

	MAT	CHE	ANT	REL	POL	ECO	GPA	μ_i
John	B-	B	B+	A-			3.18	3.51
Paul	C+	B-		B+	A-		3.00	3.16
George		C+	B-		B+	A-	3.00	2.84
Ringo			C+	B-	B	B+	2.83	2.49
Avg.	2.50	2.70	2.77	3.23	3.33	3.50		
ν_j	-0.84	-0.50	-0.18	+0.18	+0.50	+0.84		

TABLE 3

The same example as shown in Table 1 with aptitude μ_i and inflatedness ν_j shown alongside row and column grade averages.

Of course, we could minimize the sum of the absolute values instead of the sum of the squares. Generally speaking, sample means minimize the sum of squares whereas sample medians minimize the sum of absolute deviations. Medians are more robust estimators of centrality than means but it is easier to provide confidence intervals for means. For the latter reason, we will stick with summing squares for most of this paper.

Table 3 shows the output for Beatle University. The student aptitude metrics clearly show that John is the smartest Beatle. Also, while average grades in the courses correctly show that Math is the most difficult and Econ is the easiest, the inflatedness metric expands on the disparity. For example, based on averages, a student might think that the difference between Math and Econ is just one full letter grade but the inflatedness metric suggests the difference is more like one and two thirds letter grades (1.68 to be precise).

We will return to more examples later in Section 5 including one example using real-world data. But, first, let us analyze our model.

3. Least Squares. Statistical estimates of underlying unobserved fundamental quantities have little value without an associated estimate for an error in the estimation. For general least squares models, it is well understood how to produce such error bars. Nonetheless, it is instructive to derive the formulae from scratch in this particular context, at least in the particular case where we assume, unrealistically, that every student takes every course.

3.1. Estimating Means. To make a connection with utterly standard and elementary concepts, let us assume for the moment that we simply want to estimate some underlying single parameter μ based on n observations X_j , $j = 1, 2, \dots, n$. In other words, we assume that

$$X_j = \mu + \epsilon_j$$

where the ϵ_j 's are taken to be independent, identically-distributed random variables with mean zero and variance σ^2 . The parameter μ is unknown and to be estimated. The variance σ^2 is also unknown and must be estimated as well. The least squares estimator $\bar{\mu}$ for μ is that value of μ that minimizes

$$f(\mu) = \frac{1}{n} \sum_j (X_j - \mu)^2.$$

Taking the derivative and setting it equal to zero, one gets that $\bar{\mu}$ is just the sample mean:

$$\bar{\mu} = \frac{1}{n} \sum_j X_j.$$

Since the X_j 's are independent and have variance σ^2 , it follows that $\bar{\mu}$ has variance σ^2/n . The function f evaluated at $\bar{\mu}$ provides a good estimator for σ^2 :

$$\sigma^2 \approx f(\bar{\mu}) = \frac{1}{n} \sum_j \left(X_j - \frac{1}{n} \sum_k X_k \right)^2.$$

3.2. Every Student Takes Every Course. Now let's consider the problem of estimating aptitude and inflatedness from grade data. But, in an attempt to keep things simple, let us assume that every student takes every course. We have m students and n courses and therefore the set \mathcal{G} consists of mn pairs for which we have grades. As before, let f denote the function to be minimized:

$$f(\mu_1, \dots, \mu_m, \nu_1, \dots, \nu_n) = \frac{1}{mn} \sum_{i,j} (X_{ij} - \mu_i - \nu_j)^2.$$

As mentioned earlier, there is an ambiguity in the model—we could add an arbitrary constant to every aptitude and subtract that same constant from every inflatedness and the function f would be unchanged. In a previous section, we addressed this ambiguity by imposing one extra constraint, namely, that the sum of the ν_j 's be zero. We could do that here, introducing then the associated Lagrange multiplier, forming the Lagrangian, and solving the problem that way. But, it is such a simple constraint that we prefer to introduce it in a less formal manner as we go. In doing so, we hope that the analysis will be more transparent, not less.

Taking derivatives with respect to each of the variables and setting these derivatives to zero, we get the following system of equations for the estima-

tors $\bar{\mu}_j$'s and $\bar{\nu}_i$'s:

$$\begin{aligned}\bar{\mu}_i &= \frac{1}{n} \sum_j (X_{ij} - \bar{\nu}_j) \\ \bar{\nu}_j &= \frac{1}{m} \sum_i (X_{ij} - \bar{\mu}_i).\end{aligned}$$

Here, it is convenient to switch to matrix-vector notation. So, letting

$$\bar{\mu} = \begin{bmatrix} \bar{\mu}_1 \\ \bar{\mu}_2 \\ \vdots \\ \bar{\mu}_m \end{bmatrix}, \quad \bar{\nu} = \begin{bmatrix} \bar{\nu}_1 & \bar{\nu}_2 & \cdots & \bar{\nu}_n \end{bmatrix},$$

and

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix},$$

we can rewrite our optimality equations as

$$\begin{aligned}\bar{\mu} &= \frac{1}{n} (Xe - e\bar{\nu}e) \\ \bar{\nu} &= \frac{1}{m} (e^T X - e^T \bar{\mu} e^T),\end{aligned}\tag{3.1}$$

where e denotes a column vector of either m or n ones, the dimension being obvious from context. Substituting the second equation into the first, we can isolate $\bar{\mu}$:

$$\bar{\mu} = \frac{1}{n} \left(Xe - \frac{1}{m} e (e^T X - e^T \bar{\mu} e^T) e \right).$$

Collecting terms involving $\bar{\mu}$ on the left side, the remaining terms on the right-hand side, and using the fact that $e^T e = n$, we get

$$\left(I - \frac{1}{m} ee^T \right) \bar{\mu} = \left(I - \frac{1}{m} ee^T \right) \left(\frac{1}{n} Xe \right).$$

If the matrix $I - ee^T/m$ were nonsingular, we would at this point conclude that

$$\bar{\mu} = \frac{1}{n} Xe.\tag{3.2}$$

But, the matrix is singular with rank deficiency one (e is in the null space). So, there are other choices for $\bar{\mu}$. Indeed, there is a one-parameter family of choices (any $\bar{\mu}$ for which $\bar{\mu} - (1/n)Xe$ is in the null space of $I - ee^T/m$). Nonetheless, we choose to let $\bar{\mu}$ be given by (3.2) and as we shall now show this choice guarantees that the sum of the $\bar{\nu}_j$'s vanishes as we have required. Indeed, plugging (3.2) into (3.1), we get

$$(3.3) \quad \bar{\nu} = \frac{1}{m} \left(e^T X - \frac{1}{n} e^T X e e^T \right)$$

and therefore that

$$\bar{\nu} e = \frac{1}{m} \left(e^T X - \frac{1}{n} e^T X e e^T \right) e = \frac{1}{m} (e^T X e - e^T X e) = 0,$$

the second equality following from the fact that $e^T e = n$.

From (3.2) and (3.3), we see that the $\bar{\mu}_i$'s and the $\bar{\nu}_j$'s are just row and column sample means with one of them shifted by the overall mean.

Reverting back to explicit component notation, (3.2) and (3.3) can be written as

$$\begin{aligned} \bar{\mu}_i &= \frac{1}{n} \sum_j X_{ij}, & i = 1, 2, \dots, m, \\ \bar{\nu}_j &= \frac{1}{m} \sum_i X_{ij} - \frac{1}{mn} \sum_{i,j} X_{ij}, & j = 1, 2, \dots, n. \end{aligned}$$

From the first formula, we immediately see that

$$(3.4) \quad \text{var}(\bar{\mu}_j) = \frac{\sigma^2}{n}.$$

Computing the variance of the $\bar{\nu}_j$'s is a little more tedious but entirely routine. The result is

$$(3.5) \quad \text{var}(\bar{\nu}_i) = \frac{\sigma^2}{m} \left(1 - \frac{1}{n} \right) \approx \frac{\sigma^2}{m}.$$

Finally, we need an estimate of σ^2 . As before, we can use the objective function f evaluated at the optimal values for the μ_i 's and ν_j 's:

$$\sigma^2 \approx f(\bar{\mu}_1, \dots, \bar{\mu}_m, \bar{\nu}_1, \dots, \bar{\nu}_n) = \frac{1}{mn} \sum_{i,j} (X_{ij} - \bar{\mu}_i - \bar{\nu}_j)^2.$$

3.3. *Students Take Selected Courses.* Now suppose that each student takes only a small subset of the courses offered. For each student i , let $\mathcal{J}(i)$ denote the set of courses taken by student i . Similarly, for each course j , let $\mathcal{I}(j)$ denote the set of students that took course j .

The least-squares loss function is now given by

$$f(\mu_1, \dots, \mu_m, \nu_1, \dots, \nu_n) = \frac{1}{N} \sum_{(i,j) \in \mathcal{G}} (X_{ij} - \mu_i - \nu_j)^2,$$

where N denotes the cardinality of the grade-set \mathcal{G} . Again, we differentiate and set to zero. This time we get

$$(3.6) \quad \bar{\mu}_i = \frac{1}{n_i} \sum_{j \in \mathcal{J}(i)} (X_{ij} - \bar{\nu}_j) \quad i = 1, 2, \dots, m$$

$$(3.7) \quad \bar{\nu}_j = \frac{1}{m_j} \sum_{i \in \mathcal{I}(j)} (X_{ij} - \bar{\mu}_i) \quad j = 1, 2, \dots, n,$$

where n_i denotes the cardinality of $\mathcal{J}(i)$ and m_j denotes the cardinality of $\mathcal{I}(j)$. Substituting (3.7) into (3.6), we get

$$\bar{\mu}_i = \frac{1}{n_i} \sum_{j \in \mathcal{J}(i)} \left(X_{ij} - \frac{1}{m_j} \sum_{i' \in \mathcal{I}(j)} (X_{i'j} - \bar{\mu}_{i'}) \right) \quad i = 1, 2, \dots, m.$$

This is a set of m equations in m unknowns. If there is adequate diversity in student course selections so that every course indirectly is connected to every other course, then one would expect this system to have rank $m - 1$ leaving only one dimensional ambiguity in the equations. Inspired by the simplicity of the results in the previous section, we can hope that again simple sample means will provide one solution to this system of equations:

$$\bar{\mu}_i \stackrel{?}{=} \frac{1}{n_i} \sum_{j \in \mathcal{J}(i)} X_{ij} \quad i = 1, 2, \dots, m.$$

In order for this to be correct, we need to have

$$\sum_{j \in \mathcal{J}(i)} \frac{1}{m_j} \sum_{i' \in \mathcal{I}(j)} \left(X_{i'j} - \frac{1}{n_{i'}} \sum_{j' \in \mathcal{J}(i')} X_{i'j'} \right) = 0.$$

Unfortunately, there is no particular reason for this to be true. And, as we saw with the second example in the introduction, it is possible for the sample

means to be all the same even when there is a big difference in course grade inflatedness and/or in student aptitude. The model detects such differences.

Even though it appears there is no simple formula for the solution to the least-squares formulation of our problem, modern statistical and/or optimization software can solve these problems numerically without difficulty even when the data sets are very large.

Also, the fact that we have not been able to give a simple concrete formula for the $\bar{\mu}_i$'s and the $\bar{\nu}_j$'s makes it impossible to give a simple concrete formula for the variance of these random variables. Nonetheless, we can infer from the concrete results obtained before that one should first estimate σ^2 using the optimal value of the objective function as an estimate of this quantity and then the variance of the individual $\bar{\mu}_i$'s and $\bar{\nu}_j$'s can be approximated simply by dividing by the number of grades reflected in that aggregation (that is, either n_i or m_j).

4. Least Absolute Deviations. In this section, we consider a robust model in which we minimize the sum of the absolute deviations. To motivate what follows, we start with a brief review of medians.

4.1. *Medians.* As when we discussed means, let us assume for the moment that we simply want to estimate some underlying single parameter μ based on n observations X_j , $j = 1, 2, \dots, n$. In other words, we assume that

$$X_j = \mu + \epsilon_j$$

where the ϵ_j 's are taken to be independent, identically-distributed random variables with mean zero and variance σ^2 . The least absolute deviation estimator $\hat{\mu}$ for μ is the value of μ that minimizes

$$f(\mu) = \frac{1}{n} \sum_j |X_j - \mu|.$$

Taking the derivative and setting it equal to zero, one gets that $\hat{\mu}$ must satisfy

$$\sum_j \text{sgn}(X_j - \hat{\mu}) = 0,$$

which is clearly solved by setting $\hat{\mu}$ equal to the median of the X_j 's (so that half of the sgn 's are $+1$ and the other half are -1).

4.2. *Every Student Takes Every Course.* Now let's consider the problem of estimating student aptitude and course inflatedness from grade data. As

before, we start by assuming that every student takes every course. Once again, let f denote the function to be minimized:

$$f(\mu_1, \dots, \mu_m, \nu_1, \dots, \nu_n) = \frac{1}{mn} \sum_{i,j} |X_{ij} - \mu_i - \nu_j|.$$

Taking derivatives with respect to each of the variables and setting these derivatives to zero, we get the following system of equations for the estimators $\hat{\mu}_j$'s and $\hat{\nu}_i$'s:

$$\begin{aligned} \sum_j \operatorname{sgn}(X_{ij} - \hat{\mu}_i - \hat{\nu}_j) &= 0 & i = 1, 2, \dots, m \\ \sum_i \operatorname{sgn}(X_{ij} - \hat{\mu}_i - \hat{\nu}_j) &= 0 & j = 1, 2, \dots, n. \end{aligned}$$

Unlike before, there seems to be no simple description of the solution to this problem. But, we can give an algorithm that should converge quickly to the solution. Specifically, initialize

$$\begin{aligned} \hat{\nu}_j &= 0, & j = 1, 2, \dots, n \\ \hat{\mu}_i &= \operatorname{median}\{X_{ij} \mid j = 1, 2, \dots, n\}, & i = 1, 2, \dots, m. \end{aligned}$$

Then, iterate the following until there is no change from one iteration to the next:

$$\begin{aligned} \hat{\nu}_j &= \operatorname{median}\{X_{ij} - \hat{\mu}_i \mid i = 1, 2, \dots, m\}, & j = 1, 2, \dots, n \\ \hat{\mu}_i &= \operatorname{median}\{X_{ij} - \hat{\nu}_j \mid j = 1, 2, \dots, n\}, & i = 1, 2, \dots, m. \end{aligned}$$

This algorithm is unlikely to converge to a solution that satisfies $\sum_j \hat{\nu}_j = 0$ but, given the initialization, it should come close to this point.

4.3. *Students Take Selected Courses.* Finally, let us return to the general case in which each student takes only a small subset of the courses offered. The problem is to minimize the sum of the absolute values of the ϵ_{ij} 's:

$$(4.1) \quad \begin{aligned} &\text{minimize} && \sum_{(i,j) \in \mathcal{G}} |\epsilon_{ij}| \\ &\text{subject to} && X_{ij} = \mu_i + \nu_j + \epsilon_{ij} \quad \text{for } (i, j) \in \mathcal{G} \\ &&& \sum_j \nu_j = 0. \end{aligned}$$

It is easy to reformulate this model as a linear programming (LP) problem:

$$\begin{aligned} & \text{minimize} && \sum_{(i,j) \in \mathcal{G}} t_{ij} \\ & \text{subject to} && -t_{ij} \leq X_{ij} - \mu_i - \nu_j \leq t_{ij} \quad \text{for } (i,j) \in \mathcal{G} \\ & && \sum_j \nu_j = 0. \end{aligned}$$

Such linear programming problems can be solved easily. In the next section we give some examples and we compare the results from least squares formulations with those from the least absolute deviations model.

5. Examples. Finally, we consider a few specific examples including one based on real data.

5.1. *Truncated Example.* The example shown in Table 2 was contrived in order to make a point. In particular, it had an infinite number of students and courses. In Table 4, we show a truncated version consisting of eight students taking courses from a school offering eight courses. Each student takes three to five courses. As with the untruncated version, it is clear that the students are listed in order of their aptitude with the best student at the top. However, student GPA's hardly reflect the obvious trend in aptitude. The μ_i 's computed by our model make the difference in aptitude much more apparent. Similarly, average grades given in the courses show a small trend in the correct direction but they hardly account for the rather obvious overall trend in course inflatedness as one scans from left to right across the table. The ν_j 's do a much better job of identifying course inflatedness.

It is interesting to point out that the least squares and the least absolute deviation models both give the same results for this particular example.

5.2. *Circulant Example.* This example is almost the same as the truncated example in the previous subsection. Here, however, we have added two courses to Sean's schedule and to Heather's schedule and we have added one course to Yoko's schedule and to Jane's schedule. The result is a table of grades that has a *circulant* structure. Now, the trends that were clearly apparent in the truncated example are completely gone. In this example, both student GPA and the μ_i 's reflect the lack of any differentiation among the students. Similarly, the course averages and the ν_j 's both show that all courses are curved the same.

	MAT	CHE	ANT	REL	POL	ECO	HIS	SOC	GPA	μ_i
Sean	B+	A-	A						3.67	4.50
Yoko	B	B+	A-	A					3.50	4.17
John	B-	B	B+	A-	A				3.33	3.83
Paul		B-	B	B+	A-	A			3.33	3.50
George			B-	B	B+	A-	A		3.33	3.17
Ringo				B-	B	B+	A-	A	3.33	2.83
Jane					B-	B	B+	A-	3.17	2.50
Heather						B-	B	B+	3.00	2.17
Avg.	3.00	3.17	3.33	3.33	3.33	3.33	3.50	3.67		
ν_j	-1.17	-0.83	-0.50	-0.17	+0.17	+0.50	+0.83	+1.17		

TABLE 4

Truncated Example. *This is the same as the example shown in Table 2 but it has been truncated to represent a school with eight students and eight courses. Each student took three to five courses with grades as shown. As with the untruncated version, there are clear trends in student aptitude and course inflatedness, which our model correctly uncovers.*

	MAT	CHE	ANT	REL	POL	ECO	HIS	SOC	GPA	μ_i
Sean	B+	A-	A				B-	B	3.33	3.33
Yoko	B	B+	A-	A				B-	3.33	3.33
John	B-	B	B+	A-	A				3.33	3.33
Paul		B-	B	B+	A-	A			3.33	3.33
George			B-	B	B+	A-	A		3.33	3.33
Ringo				B-	B	B+	A-	A	3.33	3.33
Jane	A				B-	B	B+	A-	3.33	3.33
Heather	A-	A				B-	B	B+	3.33	3.33
Avg.	3.33	3.33	3.33	3.33	3.33	3.33	3.33	3.33		
ν_j	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		

TABLE 5

Circulant Example. *This example is the same as the previous one except that there are six more grades filling out the matrix into a circulant form. Now the trends are gone. Every student has a B+ average and every course is curved to a B+. Our model correctly assigns every course an easiness adjustment of 0.00 leaving every student's "corrected" GPA equal to his/her original GPA.*

5.3. *Test Using Simulated Data.* To confirm that the algorithm performs properly for a data set of a size and scale typical of a university, a Monte Carlo grade sample was generated. Here the degree of difficulty of the courses and the aptitudes of the simulated students are known and subject to the control of the programmer. Therefore by comparing input and output values, one can verify that basic functionality of the algorithm and can quantify its ability to determine the degree of difficulty of a course.

Monte Carlo grade samples of varying complexity were studied. In all cases, the algorithm correctly extracted the student aptitude and degree of difficulty assumed in generating the simulated data. Here we describe the results of a single model that incorporates many of the characteristics one would expect in a sample of actual grade data.

In the model to be described, there were 1000 students, each selecting four courses from a menu of 100 courses. Students were divided into five groups of different aptitudes. The average grade-point average was a 2.5. Students in the first group were assumed to have GPAs a full 1.5 letter-grade points higher than the average, students in the second had a GPA that was 0.75 grade points higher, students in the third were assumed to be average, and students in the fourth and fifth groups had averages that were 0.75 and 1.5 letter grades lower than the average.

The 100 courses were divided into five groups of 20. The relative degree of difficulty of these courses was chosen at random from a Gaussian distribution with standard deviation $\sigma = 0.5$ letter-grade points. Apart from the statistical fluctuations associated with the course-by-course randomization of degree of difficulty, the courses from the five groups had the same level of difficulty.

The assigned grades were based on the aptitude bias for the student and the course bias, both of which were subject to 0.5 letter grade Gaussian fluctuations. In addition, a particular grade for a given student in a course was subject to an additional 0.5 letter grade Gaussian fluctuation.

In the selection of courses, students in the first group chose courses disproportionately from the first group of 20 courses, students in the second group disproportionately from the second, and so on. This selection caused the average grades in the five course groups to vary (see the upper panel in Fig. 1), even though they had the same intrinsic difficulty on average.

The lower panel in Fig. 1 shows the inflatedness (easiness) parameter for the same set of courses. As one would expect for a properly functioning algorithm there is no systematic bias in this parameter, since the input model was free from any such bias. In other words, the algorithm has successfully removed the *apparent* bias in course grading resulting from the clustering of

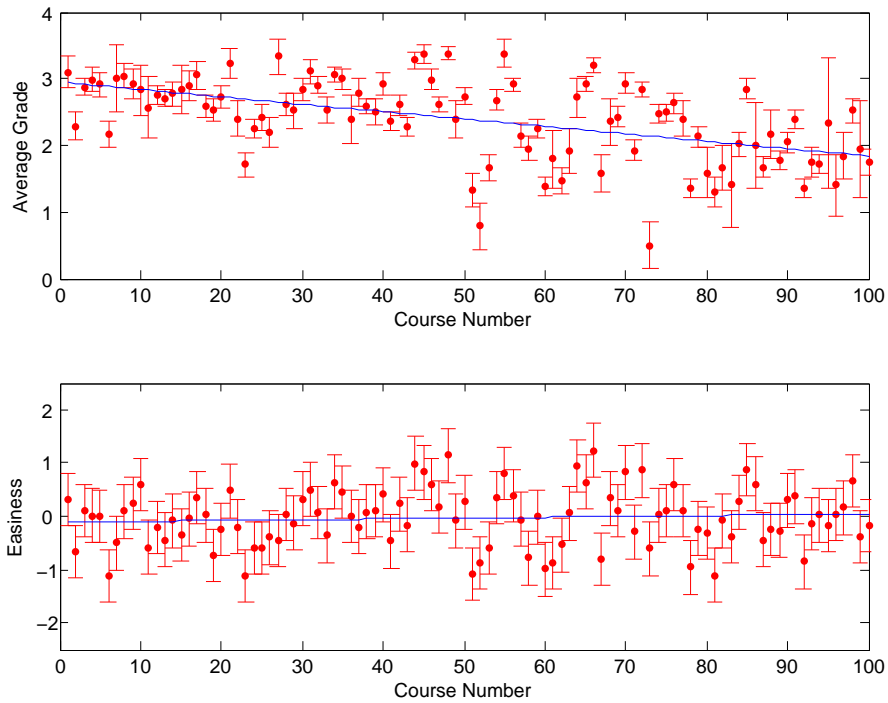


FIG 1. *Top: Average grade in each of the 100 courses. Bottom: Inflatedness parameter for the same set of courses.*

higher (lower) aptitude students in the first (last) course group.

Another way to quantify the performance of the algorithm is to histogram the difference between the average grade in each course and the degree of inflatedness that was put in to the model. The left-hand panel in Fig. 2 shows this difference. The right-hand panel in Fig. 2 shows the difference between the inflatedness calculated using the algorithm and the input inflatedness. Comparing the two histograms, one sees that the algorithm predicts the input inflatedness with an r.m.s. accuracy of 0.24 letter-grade points, which is a two-fold improvement over the simpler measure based on the average grade in the course. The residual 0.24 letter grade deviation can be attributed to statistical fluctuations in student performance.

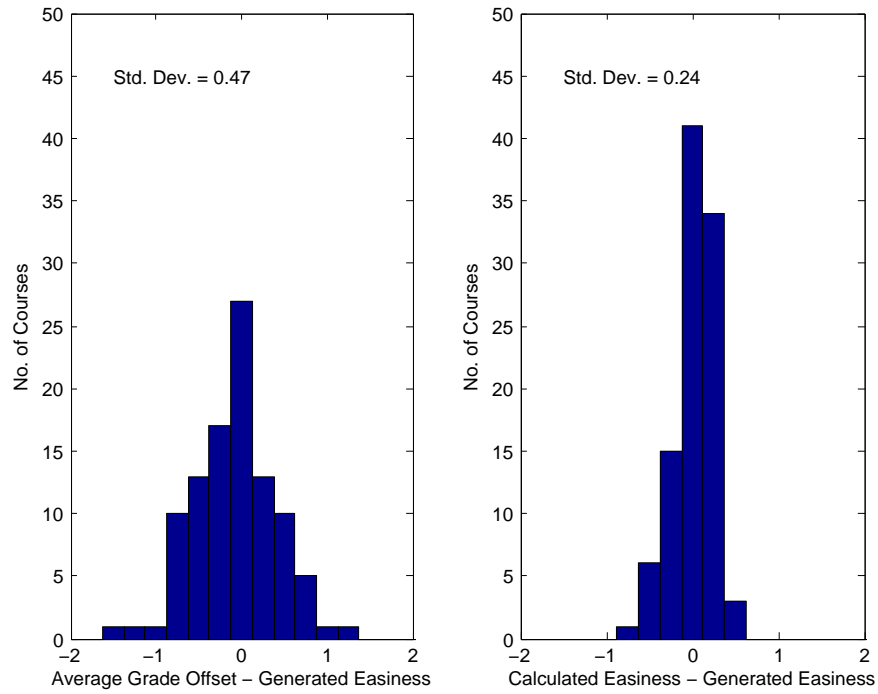


FIG 2. *Left: Average grade in each of the 100 courses. Right: Inflatedness parameter for the same set of courses.*

5.4. *Two Semesters of Real Data.* The registrar at a private university in the northeast has given us a complete two-semester data set. There are about 5000 students at this university each of whom takes four or five courses per semester from a selection of roughly 700 courses offered each semester. The data is encoded—we don’t know the identity of any particular student. Nor can we tell which course is which. All of this has been pre-encoded by the registrar. But, the grades are real. A small snippet of the data is shown in Table 6.

Table 7 shows a sample of the output from the least squares model. Table 8 shows a sample of the output from the least absolute deviations model. Comparing Tables 7 and 8, it is clear that the results are similar.

Typical courses have between 10 and 100 students. For the larger courses, there seems to be an adequate amount of data to draw conclusions. Since, the data set only represents two semesters and most students take only four or five courses in a semester, one should not put too much credence in the aptitudes assigned to the students. But, a larger data set consisting of three or four years of data would contain about 20 to 30 courses of grade data for each student. In such a case, one could imagine that the μ_i ’s would be a pretty good indicator of student aptitude.

6. Conclusions. The fundamental data available to a registrar is grading data: the X_{ij} ’s. In recent years, this data set had been used for two main purposes: (1) to assess student achievement, and (2) to assess course-by-course grade inflation. Student achievement is usually assessed by reporting on a transcript the student’s GPA. A statistical justification for this is that the totality of all student GPA’s is the simple least-squares solution to the following regression model:

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad (i, j) \in \mathcal{G}.$$

At the same time, grade inflation is assessed by reporting average (or median) grades given in a course. The totality of average course grades is the least-squares solution to a “dual” regression model:

$$X_{ij} = \nu_j + \epsilon_{ij}, \quad (i, j) \in \mathcal{G}.$$

It seems only natural that these two problems should be combined into one and that is exactly what we have proposed in this paper.

Grade inflation, and what to do about it, has been discussed extensively in recent years. In this paper, we have described an analytical approach to disentangling the course-by-course differences in grading policies from underlying student aptitudes. If such a tool were to be widely adopted and

1	F001090	3.7	5	F002046	3	8	S008811	2.7
1	F004148	1.7	5	F005976	3.7	8	S010952	3.7
1	F006665	2	5	F007285	2.3	8	S010973	1.7
1	F010449	3	5	F008991	4	9	F002614	2.7
1	S009167	3.7	5	F010762	4	9	F006664	2
1	S009571	2	5	S001380	3.7	9	F008144	1.7
1	S010994	2.7	5	S004153	2.3	9	F008832	2.7
2	F003387	3	5	S005842	4	9	F010542	3
2	F009193	2.7	5	S008310	4	9	S001065	3.3
2	F010693	3	6	F001400	2.7	9	S001542	2
2	F010813	2.7	6	F004647	3.7	9	S004398	2.3
2	S001093	1	6	F006787	3.3	9	S004399	2.3
2	S003408	3	6	F009999	2.7	10	F008991	2
2	S005302	2	6	S003424	3	10	F009582	2.3
3	F003769	3	6	S003952	3	10	S001404	4
3	F004893	3.7	6	S009187	3.7	10	S002463	4
3	F004896	3.3	6	S010953	3.7	10	S004186	4
3	S004172	1.7	7	F005979	3.3	10	S004398	2
4	F000613	3.7	7	F007230	3.3	11	F001090	3.3
4	F001381	2.7	7	F010437	3.3	11	F001109	3.7
4	F004140	3	7	S006804	4	11	F003243	1.7
4	F005588	3.7	7	S010960	4	11	F005558	2.3
4	S000185	3	8	F001064	4	11	S002625	2.3
4	S004398	3	8	F002461	2.3	11	S007854	2.7
4	S004901	2.7	8	F005979	3	11	S010979	3.3
4	S009698	3.3	8	S007946	0		:	

TABLE 6

One semester of data consisting of about 37000 grades given to about 5000 students. Each record consists of three data elements: the student id (encoded), the course id (also encoded), and the grade (converted from a letter grade to a numerical grade in the usual manner).

student aptitude as defined by the models given in this paper were to become the accepted measure of student accomplishment, then issue of standardizing grading policies across a university becomes somewhat moot.

Of course, there is still the important question of comparing grades from students across different universities, which is something professional schools, graduate schools, and employers must do routinely. Unfortunately, the model described here cannot address this difficult problem without a dataset in which students at divergent universities take common courses. Perhaps the only way to do that would be to make a huge model in which all high-school and university grading data are fed into one huge master program. If such data were ever made available, which is highly doubtful, such a problem would probably be too large to solve on today's computers.

Acknowledgement. The authors would like to thank Jianqing Fan for useful discussions regarding underlying statistical ideas.

F001204	-2.55 ± 0.50	1	S008128	-0.54 ± 0.36	2	F010864	0.64 ± 0.29	3
F002509	-2.49 ± 0.50	1	F002339	-0.53 ± 0.11	22	S002603	0.64 ± 0.13	14
S001225	-1.77 ± 0.50	1	F004137	-0.53 ± 0.10	25	S008485	0.64 ± 0.12	17
S003935	-1.04 ± 0.36	2	F008314	-0.53 ± 0.15	11	F000295	0.65 ± 0.13	14
F003936	-0.89 ± 0.50	1	F009959	-0.53 ± 0.13	14	F010480	0.65 ± 0.25	4
S002963	-0.86 ± 0.09	33	F010275	-0.52 ± 0.29	3	S010396	0.66 ± 0.29	3
S005818	-0.77 ± 0.17	9	S008328	-0.52 ± 0.15	11	F010501	0.68 ± 0.50	1
S004319	-0.75 ± 0.23	5	F005558	-0.51 ± 0.04	187	F002968	0.69 ± 0.50	1
S008329	-0.70 ± 0.15	12	S000519	-0.51 ± 0.15	12	F009955	0.69 ± 0.36	2
S003007	-0.68 ± 0.21	6	S001093	-0.50 ± 0.07	47	S007268	0.69 ± 0.17	9
S001783	-0.66 ± 0.08	36	S008624	-0.50 ± 0.19	7	S010988	0.73 ± 0.18	8
S010294	-0.66 ± 0.29	3	F001003	-0.49 ± 0.19	7	F010535	0.74 ± 0.50	1
F004151	-0.65 ± 0.05	107	F002060	-0.49 ± 0.12	18	F010783	0.75 ± 0.36	2
F008345	-0.60 ± 0.15	11		S008506	0.78 ± 0.29	3
S002477	-0.60 ± 0.17	9	S001543	0.58 ± 0.16	10	S009990	0.78 ± 0.29	3
S004159	-0.60 ± 0.08	42	S007263	0.58 ± 0.23	5	S010720	0.78 ± 0.29	3
F004140	-0.59 ± 0.05	122	S010725	0.58 ± 0.19	7	S010987	0.80 ± 0.18	8
F008328	-0.59 ± 0.16	10	S010932	0.58 ± 0.29	3	F010617	0.81 ± 0.21	6
S001380	-0.59 ± 0.03	312	F010402	0.59 ± 0.18	8	F010830	0.84 ± 0.29	3
F004153	-0.58 ± 0.15	11	S004063	0.60 ± 0.23	5	S010986	0.90 ± 0.50	1
S009395	-0.58 ± 0.29	3	S004870	0.60 ± 0.25	4	S000205	0.93 ± 0.50	1
F009200	-0.57 ± 0.18	8	F005922	0.61 ± 0.21	6	S011047	0.96 ± 0.29	3
F010277	-0.57 ± 0.25	4	F010395	0.61 ± 0.29	3	S010039	1.06 ± 0.50	1
F004322	-0.56 ± 0.07	55	F004189	0.62 ± 0.25	4	F003038	1.22 ± 0.50	1
F005128	-0.55 ± 0.03	256	S001263	0.62 ± 0.18	8	S010261	1.66 ± 0.36	2
S004150	-0.55 ± 0.03	217	F004043	0.63 ± 0.21	6	F010122	1.92 ± 0.50	1

TABLE 7

A partial listing of the course inflatedness associated with the data partially shown in Table 6. The table shows in three columns the beginning and the end of a long table of data with three columns. The first column is the course id, the second column is the inflatedness ν_j , and the third column shows the course enrollment. In the interest of space, we show only some of the least inflated courses and some of the most inflated courses. It is interesting to note that, with the exception of a few very small classes (seminar and project courses), the inflatedness spans from about -0.45 to 0.55 . In other words, a student can expect a plus/minus half-letter grade deviation from his/her "true" aptitude simply because of differences in grading policies among some courses.

F001204	-3.22 ± 0.36	1	F001392	-0.46 ± 0.06	38	S004206	0.64 ± 0.18	4
F002509	-2.49 ± 0.36	1	F001403	-0.46 ± 0.03	171	S005917	0.64 ± 0.11	10
S001225	-1.78 ± 0.36	1	F001759	-0.46 ± 0.06	37	F005099	0.67 ± 0.15	6
S003935	-1.30 ± 0.25	2	F002376	-0.46 ± 0.25	2	S003046	0.67 ± 0.18	4
F010315	-1.06 ± 0.21	3	F002969	-0.46 ± 0.08	19	S010342	0.70 ± 0.25	2
F003936	-0.87 ± 0.36	1	F004140	-0.46 ± 0.03	122	F004043	0.71 ± 0.15	6
S002491	-0.82 ± 0.21	3	F004148	-0.46 ± 0.06	40	F000295	0.74 ± 0.10	14
S002963	-0.76 ± 0.06	33	F004149	-0.46 ± 0.03	188	F004189	0.74 ± 0.18	4
S008128	-0.70 ± 0.25	2	F004150	-0.46 ± 0.04	96	F010783	0.74 ± 0.25	2
F004151	-0.66 ± 0.03	107	F004408	-0.46 ± 0.08	18	S007263	0.74 ± 0.16	5
F008328	-0.66 ± 0.11	10	F005128	-0.46 ± 0.02	256	S009571	0.74 ± 0.21	3
F010275	-0.66 ± 0.21	3	F005558	-0.46 ± 0.03	187	F010830	0.79 ± 0.21	3
S005818	-0.66 ± 0.12	9	F006660	-0.46 ± 0.05	45	S010986	0.80 ± 0.36	1
F010277	-0.60 ± 0.18	4		F010535	0.83 ± 0.36	1
F009049	-0.59 ± 0.11	11	S010987	0.62 ± 0.13	8	F001267	0.84 ± 0.36	1
F009959	-0.58 ± 0.10	14	F008752	0.63 ± 0.21	3	F003038	0.84 ± 0.36	1
F001003	-0.56 ± 0.14	7	F002968	0.64 ± 0.36	1	F008385	0.84 ± 0.14	7
F004322	-0.56 ± 0.05	55	F004923	0.64 ± 0.14	7	S000205	0.84 ± 0.36	1
S001783	-0.56 ± 0.06	36	F010501	0.64 ± 0.36	1	S004870	0.84 ± 0.18	4
S004159	-0.56 ± 0.06	42	F010617	0.64 ± 0.15	6	S010396	0.84 ± 0.21	3
S005334	-0.56 ± 0.14	7	S000522	0.64 ± 0.21	3	S011047	0.84 ± 0.21	3
S008329	-0.56 ± 0.10	12	S001543	0.64 ± 0.11	10	S008506	0.88 ± 0.21	3
S008344	-0.56 ± 0.10	12	S001550	0.64 ± 0.11	10	S009990	0.88 ± 0.21	3
F004180	-0.52 ± 0.09	16	S002603	0.64 ± 0.10	14	S010720	0.88 ± 0.21	3
F009519	-0.51 ± 0.15	6	S003304	0.64 ± 0.16	5	S010924	0.94 ± 0.16	5
S008328	-0.51 ± 0.11	11	S004063	0.64 ± 0.16	5	S010261	1.08 ± 0.25	2

TABLE 8

A partial listing of the course inflatedness associated with the data partially shown in Table 6 as computed using the least absolute deviations model.

DEPT. OF OPERATIONS RESEARCH AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON NJ 08544
E-MAIL: rvdb@princeton.edu
gordon.scharf@gmail.com
marlow@princeton.edu