Abstract. In this paper, we consider an issue of great interest to all students: fairness in grading. Specifically, we represent each grade as a student’s intrinsic (overall) aptitude minus a correction representing the course’s inherent difficulty plus a statistical error. We consider two statistical methods for assigning an aptitude to each student and, simultaneously, a measure of difficulty to each course: (1) we minimize the sum of squares of the errors and (2) we minimize the sum of the absolute values of those errors. We argue that by accounting for course difficulty, we arrive at a measure of aptitude that is fairer than the usual grade-point-average metric. At the same time, the measures of course difficulty can be used to inform instructors as to how their courses compare to others. The two particular models presented are examples of least-squares and least-absolute-deviation regression and can be used in the classroom to motivate an interest in regression in general and to illustrate the pros and cons of these two approaches to the regression problem.

AMS subject classifications. 62J05, 62P25

Key words. mean, median, least squares, least absolute deviations

1. Introduction. Course assessment and grading policy are topics of great interest to most students. Mathematical models that address inherent unfairness in the assessment process provide an excellent example of regression that can be taught in undergraduate statistics and/or optimization courses. In fact, one of us (Scharf) was a junior contemplating what would make an interesting senior thesis and after a casual dinner conversation with classmates came up with the idea that a statistical method to adjust student grade-point averages according to the difficulty of the courses taken could lead to a very interesting thesis. This article, while it highlights a different statistical approach than the one originally proposed, is an outgrowth of that thesis.

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 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 1.1. From this table, we see that George and Paul have received the same grades (in different courses) and so their grade-point averages (GPA’s) are the same. Furthermore, John’s grades are only slightly better and Ringo’s 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.

---

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (rvdb@princeton.edu).
†gordon.scharf@gmail.com.
‡Department of Physics, Princeton University, Princeton, NJ 08544 (marlow@princeton.edu).
### Table 1.1
Grading data from Beatle University. The six courses are Math (MAT), Chemical Engineering (CHE), Anthropology (ANT), Religion (REL), Politics (POL), and Economics (ECO).

<table>
<thead>
<tr>
<th></th>
<th>MAT</th>
<th>CHE</th>
<th>ANT</th>
<th>REL</th>
<th>POL</th>
<th>ECO</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paul</td>
<td>C+</td>
<td>B−</td>
<td>B+</td>
<td>A−</td>
<td></td>
<td></td>
</tr>
<tr>
<td>George</td>
<td>C+</td>
<td>B−</td>
<td>B+</td>
<td>A−</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ringo</td>
<td>C+</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td></td>
</tr>
</tbody>
</table>

### Table 1.2
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.

<table>
<thead>
<tr>
<th></th>
<th>MAT</th>
<th>CHE</th>
<th>ANT</th>
<th>REL</th>
<th>POL</th>
<th>ECO</th>
<th>HIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paul</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>George</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ringo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hence, GPA does not tell an unbiased 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 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 1.1, 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 1.1 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 1.2 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. The 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 takes only a small sample of the complete suite of courses offered.

We assume that each student has an aptitude\(^1\) \(\mu_i, i = 1, 2, \ldots, m\), which is unknown to us and which we wish to estimate, and that each course has an inflatedness \(\nu_j, j = 1, 2, \ldots, n\), 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:

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

(2.1)

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 \(\epsilon_{ij}\)’s are the “errors” one needs to add to make the approximation an equality. These errors reflect both the randomness associated with how any student might perform in any particular course and also a systematic deviation between the student’s overall aptitude and his/her subject-specific aptitude for the material in the particular course.

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

\[
\sum_j \nu_j = 0.
\]

(2.2)

This constraint is, of course, by choice. We need some sort of normalization. Without one, we could add an arbitrary constant to every \(\mu_i\) and subtract the same constant from every \(\nu_j\) without changing any of the \(\epsilon_{ij}\)’s.

Our aim is to find the best “fit” to the data. That is, we wish to choose the \(\mu_i\)’s and the \(\nu_j\)’s in such a manner as to make the \(\epsilon_{ij}\)’s as small as possible. To do this, we minimize the sum of the squares of the \(\epsilon_{ij}\)’s:

\[
\text{minimize} \quad \sum_{(i,j) \in \mathcal{G}} \epsilon_{ij}^2
\]

subject to

\[
X_{ij} = \mu_i + \nu_j + \epsilon_{ij} \quad \text{for } (i, j) \in \mathcal{G}
\]

\[
\sum_j \nu_j = 0.
\]

(2.3)

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 2.1 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

\(^1\)Several colleagues have pointed out the obvious fact that aptitude varies from subject to subject. We are not trying to capture this variation. In this paper, we consider “aptitude” to be a synonym for “modified GPA”—a one-dimensional parameter that could be used to determine class rank, awards, etc.
Table 2.1
The same example as shown in Table 1.1 with aptitude $\mu_i$ and inflatedness $\nu_j$ shown alongside row and column grade averages.

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. We will also discuss briefly the important question of grade-compression, which often goes hand-in-hand with grade inflation. But, first, let us analyze the model given by (2.3).

3. Least Squares. We start by giving a thorough analysis of the special case where we assume every student takes every course. After that, we generalize to the more realistic scenario in which each student takes only a small sample of the entire suite of courses.

3.1. Every Student Takes Every Course. We have $m$ students and $n$ courses and therefore, assuming that every student takes every course, the set $G$ consists of $mn$ pairs for which we have grades. Eliminating the $\epsilon_{i,j}$’s from (2.3), let $f$ denote the function to be minimized:

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

(Note that we have divided by $mn$ to make this be an average squared error instead of a total squared error.) 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 $\nu_j$’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 statistical estimators—the $\bar{\mu}_j$’s and $\bar{\nu}_i$’s—of the underlying unknown parameters (the $\mu_j$’s and the $\nu_i$’s with-

<table>
<thead>
<tr>
<th></th>
<th>MAT</th>
<th>CHE</th>
<th>ANT</th>
<th>REL</th>
<th>POL</th>
<th>ECO</th>
<th>GPA</th>
<th>$\mu_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td></td>
<td></td>
<td>3.18</td>
<td>3.51</td>
</tr>
<tr>
<td>Paul</td>
<td>C+</td>
<td>B−</td>
<td>B+</td>
<td>A−</td>
<td></td>
<td></td>
<td>3.00</td>
<td>3.16</td>
</tr>
<tr>
<td>George</td>
<td>C+</td>
<td>B−</td>
<td>B+</td>
<td>A−</td>
<td></td>
<td></td>
<td>3.00</td>
<td>2.84</td>
</tr>
<tr>
<td>Ringo</td>
<td>C+</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td></td>
<td></td>
<td>2.83</td>
<td>2.49</td>
</tr>
<tr>
<td>Avg.</td>
<td>2.50</td>
<td>2.70</td>
<td>2.77</td>
<td>3.23</td>
<td>3.33</td>
<td>3.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\nu_j$</td>
<td>−0.84</td>
<td>−0.50</td>
<td>−0.18</td>
<td>+0.18</td>
<td>+0.50</td>
<td>+0.84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
out the bars):

\[
\bar{\mu}_i = \frac{1}{n} \sum_j (X_{ij} - \bar{\nu}_j)
\]

\[
\bar{\nu}_j = \frac{1}{m} \sum_i (X_{ij} - \bar{\mu}_i).
\]

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

\[
\bar{\mu} = \begin{bmatrix}
\bar{\mu}_1 \\
\bar{\mu}_2 \\
\vdots \\
\bar{\mu}_m
\end{bmatrix}, \quad \bar{\nu} = \begin{bmatrix}
\bar{\nu}_1 \\
\bar{\nu}_2 \\
\vdots \\
\bar{\nu}_n
\end{bmatrix},
\]

and

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

we can rewrite our optimality equations as

\[
\bar{\mu} = \frac{1}{n} (Xe - e\bar{\nu}e)
\]

(3.1)

\[
\bar{\nu} = \frac{1}{m} (e^T X - e^T \bar{\mu}e^T),
\]

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

(3.2)

\[
\bar{\mu} = \frac{1}{n} Xe.
\]

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)

\[
\bar{\nu} = \frac{1}{m} \left( e^T X - \frac{1}{n} e^T X ee^T \right).
\]
and therefore that
\[ \bar{\nu} = \frac{1}{m} \left( e^T X - \frac{1}{n} e^T X e e^T \right) e = \frac{1}{m} \left( e^T X e - e^T X e \right) = 0, \]
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
\[ \bar{\mu}_i = \frac{1}{n} \sum_j X_{ij}, \quad i = 1, 2, \ldots, m, \]
\[ \bar{\nu}_j = \frac{1}{m} \sum_i X_{ij} - \frac{1}{mn} \sum_{i,j} X_{ij}, \quad j = 1, 2, \ldots, n. \]
From the first formula, we immediately see that
\[ \text{var}(\bar{\mu}_j) = \frac{\sigma^2}{n}. \quad (3.4) \]
Computing the variance of the \( \bar{\nu}_j \)'s is only slightly more tedious:
\[ \text{var}(\bar{\nu}_i) = \frac{\sigma^2}{m} \left( 1 - \frac{1}{n} \right) \approx \frac{\sigma^2}{m}. \quad (3.5) \]
Finally, we need an estimate of \( \sigma^2 \). We can use the objective function \( f \) evaluated at the optimal values for the \( \mu_i \)'s and \( \nu_j \)'s:
\[ \sigma^2 \approx f(\bar{\mu}_1, \ldots, \bar{\mu}_m, \bar{\nu}_1, \ldots, \bar{\nu}_n) = \frac{1}{mn} \sum_{i,j} (X_{ij} - \bar{\mu}_i - \bar{\nu}_j)^2. \]

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

The least-squares loss function is now given by
\[ f(\mu_1, \ldots, \mu_m, \nu_1, \ldots, \nu_n) = \frac{1}{N} \sum_{(i,j) \in G} (X_{ij} - \mu_i - \nu_j)^2, \]
where \( N \) denotes the cardinality of the grade-set \( G \). Again, we differentiate and set to zero. This time we get
\[ \bar{\mu}_i = \frac{1}{n_i} \sum_{j \in J(i)} (X_{ij} - \bar{\nu}_j) \quad i = 1, 2, \ldots, m \quad (3.6) \]
\[ \bar{\nu}_j = \frac{1}{m_j} \sum_{i \in I(j)} (X_{ij} - \bar{\mu}_i) \quad j = 1, 2, \ldots, n, \quad (3.7) \]
where \( n_i \) denotes the cardinality of \( J(i) \) and \( m_j \) denotes the cardinality of \( I(j) \).
Substituting (3.7) into (3.6), we get
\[ \bar{\mu}_i = \frac{1}{n_i} \sum_{j \in J(i)} \left( X_{ij} - \frac{1}{m_j} \sum_{' \in I(j)} (X'_{ij} - \bar{\nu}) \right) \quad i = 1, 2, \ldots, m. \]
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 subsection, we could hope that again simple sample means might provide one solution to this system of equations:

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

But, in order for this to be correct, we need to have

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

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 there is no simple formula for the solution to the least-squares formulation of our problem, it is a linear system and therefore can be solved numerically using Gaussian elimination or some fancier algorithm for solving linear equations. Alternatively, a simple iterative scheme, called the method of successive approximations, can be used to solve these equations. Start by letting the $\bar{\nu}_j$’s be all zero and use (3.6) to compute the $\bar{\mu}_i$’s. Then, use these values for the $\bar{\mu}_i$’s in (3.7) to compute updated estimates for the $\bar{\nu}_j$’s. Substitute these back into (3.6), etc. If the linear operator defining this iterative sequence is a contraction operator (and it is!), then the sequence of $\bar{\mu}_i$’s and $\bar{\nu}_j$’s will converge to their correct limiting values.

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 $\sigma^2$ 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.

4.1. Every Student Takes Every Course. As before, we start by assuming that every student takes every course. Once again, let $f$ denote the function to be minimized:

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

Note that the absolute-value function is convex but not smooth as it is not differentiable at zero. Convex functions, even non-smooth ones, have well-defined subgradients, which means that the “derivative” of a nonsmooth convex function is a multi-valued function (see, e.g., [7] or [2] for a thorough treatment of these issues). In the case of the absolute-value function, its “derivative” at zero can be taken as all
real numbers between $-1$ and $1$. With this subgradient notion of derivative, we can take derivatives with respect to each of the variables and set these derivatives to zero. We get the following system of equations for the statistical estimators—the $\hat{\mu}_j$’s and $\hat{\nu}_i$’s—of the underlying unknown parameters (same variables without the hats):

$$
\sum_j \text{sgn}(X_{ij} - \hat{\mu}_i - \hat{\nu}_j) = 0 \quad i = 1, 2, \ldots, m
$$

$$
\sum_i \text{sgn}(X_{ij} - \hat{\mu}_i - \hat{\nu}_j) = 0 \quad j = 1, 2, \ldots, n.
$$

Unlike before, there seems to be no simple description of the solution to this problem. But, as before, we can use the method of successive approximation to come up with an algorithm that should converge quickly to the solution. Specifically, initialize $\hat{\nu}_j = 0$, $j = 1, 2, \ldots, n$

$\hat{\mu}_i = \text{median}\{X_{ij} \mid j = 1, 2, \ldots, n\}$, $i = 1, 2, \ldots, m$.

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

$\hat{\nu}_j = \text{median}\{X_{ij} - \hat{\mu}_i \mid i = 1, 2, \ldots, m\}$, $j = 1, 2, \ldots, n$

$\hat{\mu}_i = \text{median}\{X_{ij} - \hat{\nu}_j \mid j = 1, 2, \ldots, n\}$, $i = 1, 2, \ldots, m$.

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. Furthermore, an appropriate shift can be applied after the algorithm converges.

**4.2. 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 least absolute deviations problem, then, is to minimize

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

$$
= \frac{1}{N} \sum_i \sum_{j \in J(i)} |X_{ij} - \mu_i - \nu_j|
$$

$$
= \frac{1}{N} \sum_j \sum_{i \in I(j)} |X_{ij} - \mu_i - \nu_j|.
$$

Again, we can equate derivatives to zero to get conditions that define the $\hat{\mu}_i$’s and the $\hat{\nu}_j$’s:

$$
\sum_{j \in J(i)} \text{sgn}(X_{ij} - \hat{\mu}_i - \hat{\nu}_j) = 0 \quad i = 1, 2, \ldots, m
$$

$$
\sum_{i \in I(j)} \text{sgn}(X_{ij} - \hat{\mu}_i - \hat{\nu}_j) = 0 \quad j = 1, 2, \ldots, n.
$$

In this case, the method of successive approximation is started by setting

$\hat{\nu}_j = 0$, $j = 1, 2, \ldots, n$

$\hat{\mu}_i = \text{median}\{X_{ij} \mid j \in J(i)\}$, $i = 1, 2, \ldots, m$. 

and iterates
\[\hat{\nu}_j = \text{median}\{X_{ij} - \hat{\mu}_i \mid i \in I(j)\}, \quad j = 1, 2, \ldots, n\]
\[\hat{\mu}_i = \text{median}\{X_{ij} - \hat{\nu}_j \mid j \in J(i)\}, \quad i = 1, 2, \ldots, m\]
until all values remain fixed for one full iteration.

Alternatively, the problem can be formulated as minimizing the sum of the absolute values of the \(\epsilon_{ij}\)'s:

\[
\begin{align*}
\text{minimize} & \quad \sum_{(i,j) \in G} |\epsilon_{ij}| \\
\text{subject to} & \quad X_{ij} = \mu_i + \nu_j + \epsilon_{ij} \quad \text{for } (i,j) \in G \\
& \quad \sum_j \nu_j = 0.
\end{align*}
\]

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

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

Such linear programming problems can be solved quickly using widely available software.

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. We consider a few specific examples including one based on real data.

5.1. Truncated Example. The example shown in Table 1.2 was contrived in order to make a point. In particular, it had an infinite number of students and courses. In Table 5.1, 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 \(\mu_i\)’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 \(\nu_j\)’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
### Table 5.1
Truncated Example. This is the same as the example shown in Table 1.2 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.

<table>
<thead>
<tr>
<th></th>
<th>MAT</th>
<th>CHE</th>
<th>ANT</th>
<th>REL</th>
<th>POL</th>
<th>ECO</th>
<th>HIS</th>
<th>SOC</th>
<th>GPA</th>
<th>$\mu_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sean</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.67</td>
<td>4.50</td>
</tr>
<tr>
<td>Yoko</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.50</td>
<td>4.17</td>
</tr>
<tr>
<td>John</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.83</td>
</tr>
<tr>
<td>Paul</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.50</td>
</tr>
<tr>
<td>George</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.17</td>
</tr>
<tr>
<td>Ringo</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>2.83</td>
</tr>
<tr>
<td>Jane</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>2.50</td>
</tr>
<tr>
<td>Heather</td>
<td></td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>2.00</td>
</tr>
<tr>
<td>Avg.</td>
<td>3.00</td>
<td>3.17</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.67</td>
<td>4.17</td>
</tr>
<tr>
<td>$\nu_j$</td>
<td>-1.17</td>
<td>-0.83</td>
<td>-0.50</td>
<td>-0.17</td>
<td>+0.17</td>
<td>+0.50</td>
<td>+0.83</td>
<td>+1.17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 5.2
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.

<table>
<thead>
<tr>
<th></th>
<th>MAT</th>
<th>CHE</th>
<th>ANT</th>
<th>REL</th>
<th>POL</th>
<th>ECO</th>
<th>HIS</th>
<th>SOC</th>
<th>GPA</th>
<th>$\mu_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sean</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>Yoko</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>John</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>Paul</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>George</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>Ringo</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>Jane</td>
<td>A</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td>A−</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>Heather</td>
<td>A−</td>
<td>A</td>
<td>B−</td>
<td>B</td>
<td>B+</td>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>Avg.</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
<td>3.33</td>
</tr>
<tr>
<td>$\nu_j$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

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. 5.1), even though they had the same intrinsic difficulty on average.

The lower panel in Fig. 5.1 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 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. 5.2 shows this difference. The right-hand panel in Fig. 5.2 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.

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 5.3.

Table 5.4 shows a sample of the output from the least squares model. Table 5.5 shows a sample of the output from the least absolute deviations model. Comparing
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 $\mu_i$’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 has 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 student aptitude as defined by the models given in this paper were to become the accepted measure of student accomplishment, then the issue of standardizing grading policies across a university becomes somewhat moot.

An issue closely related to grade inflation is grade compression: as grades have gone up, the standard deviation of the grades given has gone down. Adjusting grades for “course difficulty” as described in the paper is a good thing. An even better adjustment would also normalize grades in such a way that the standard deviation of the renormalized grades in each course is roughly uniform from one course to the next. However, this can be a tricky back-end adjustment because any small class in which all students were given the same grade would be “unspreadable”. And every school has some such courses. Anyway, we leave the resolution of this issue as a question for future research.

Finally, 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 uni-
versities 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 might prove too large to solve on today’s computers.

The models presented in this paper are good examples of least-squares and least-absolute deviations regression and can therefore be used as a pedagogical tool when teaching these topics in statistics and/or optimization courses.


Acknowledgement. The authors would like to thank Jianqing Fan for useful discussions regarding underlying statistical ideas.
A partial listing of the course inflatedness associated with the data partially shown in Table 5.3. 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 $\nu_j$, 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.

<table>
<thead>
<tr>
<th>Course ID</th>
<th>Inflatedness</th>
<th>Enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td>S001204</td>
<td>$-2.55 \pm 0.50$</td>
<td>1</td>
</tr>
<tr>
<td>F002509</td>
<td>$-2.49 \pm 0.50$</td>
<td>1</td>
</tr>
<tr>
<td>S002255</td>
<td>$-1.77 \pm 0.50$</td>
<td>1</td>
</tr>
<tr>
<td>S003935</td>
<td>$1.04 \pm 0.36$</td>
<td>2</td>
</tr>
<tr>
<td>S003936</td>
<td>$-0.89 \pm 0.50$</td>
<td>1</td>
</tr>
<tr>
<td>S002963</td>
<td>$-0.86 \pm 0.09$</td>
<td>33</td>
</tr>
<tr>
<td>S005858</td>
<td>$-0.77 \pm 0.17$</td>
<td>9</td>
</tr>
<tr>
<td>S004319</td>
<td>$-0.75 \pm 0.23$</td>
<td>5</td>
</tr>
<tr>
<td>S008329</td>
<td>$-0.70 \pm 0.15$</td>
<td>12</td>
</tr>
<tr>
<td>S003007</td>
<td>$-0.68 \pm 0.21$</td>
<td>6</td>
</tr>
<tr>
<td>S001783</td>
<td>$-0.66 \pm 0.08$</td>
<td>36</td>
</tr>
<tr>
<td>S010294</td>
<td>$-0.66 \pm 0.29$</td>
<td>3</td>
</tr>
<tr>
<td>F004151</td>
<td>$-0.65 \pm 0.05$</td>
<td>107</td>
</tr>
<tr>
<td>F008345</td>
<td>$-0.60 \pm 0.15$</td>
<td>11</td>
</tr>
<tr>
<td>S002477</td>
<td>$-0.60 \pm 0.17$</td>
<td>9</td>
</tr>
<tr>
<td>S004159</td>
<td>$-0.60 \pm 0.08$</td>
<td>42</td>
</tr>
<tr>
<td>F004140</td>
<td>$-0.59 \pm 0.05$</td>
<td>122</td>
</tr>
<tr>
<td>F008328</td>
<td>$-0.59 \pm 0.16$</td>
<td>10</td>
</tr>
<tr>
<td>S001380</td>
<td>$-0.59 \pm 0.03$</td>
<td>312</td>
</tr>
<tr>
<td>F004153</td>
<td>$-0.58 \pm 0.15$</td>
<td>11</td>
</tr>
<tr>
<td>S009395</td>
<td>$-0.58 \pm 0.29$</td>
<td>3</td>
</tr>
<tr>
<td>F009200</td>
<td>$-0.57 \pm 0.18$</td>
<td>8</td>
</tr>
<tr>
<td>F010277</td>
<td>$-0.57 \pm 0.25$</td>
<td>4</td>
</tr>
<tr>
<td>F004322</td>
<td>$-0.56 \pm 0.07$</td>
<td>55</td>
</tr>
<tr>
<td>F005128</td>
<td>$-0.55 \pm 0.03$</td>
<td>256</td>
</tr>
<tr>
<td>S001450</td>
<td>$-0.55 \pm 0.03$</td>
<td>217</td>
</tr>
</tbody>
</table>

REFERENCES

A partial listing of the course inflatedness associated with the data partially shown in Table 5.3 as computed using the least absolute deviations model.

| 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 | F004418 | -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 |