A CRITICAL DISCUSSION OF GAP ANALYSIS

JAN DE LEEUW

Contents

1. Introduction

In this report we shall present a statistical analysis of the regression, or *gap analysis*, in Dr. Armor's report. Our discussion is based on the report itself, on the Excel data files used by Dr. Armor, and on the output files of the regression analyses. We have also used previous testimony by Dr. Armor in *Coalition to Save Our Children v. Delaware State Board of Education* and *Swann v. Charlotte-Mecklenburg Board of Education*.

2. Gap Analysis

Gap analysis is not a standard or particularly well-known data analysis procedure, but there are many variations of it in the social and educational research literature. As we shall argue below, all these variations share the same fundamental flaws.

Let us give a brief outline first. There is a binary variable *race* in the regression analysis indicating if a student is black or not. We first do a regression analysis using only the *race* variable (and a constant). This gives a regression coefficient for *race*, say β_0, which is simply the difference in mean test score of blacks and non-blacks. This difference is called the *raw gap* (for this particular test and this particular group of students).

The next step in gap analysis is to attempt to "correct for background". We do a second regression analysis on the same data, where we now use the variable *race* together with some other variables we want to "correct" for. These are often called *covariates* or *background*. In Dr. Armor's case these are *lunch* and early testscores (either from first grade or kindergarten). Each of these variables gets a regression coefficient. Say the regression coefficient for *race* in this second regression analysis is β_1. This is the *corrected gap*.

Finally, Dr. Armor argues that the background "explains"

$$\left\{\frac{\beta_0-\beta_1}{\beta_0}\right\}\times 100\%$$

of the raw gap. Thus (see below) if $\beta_0 = 16.28$ and $\beta_1 = 9.26$ (as we have for reading in grade 8), then the background "explains"

$$\left\{\frac{16.28-9.26}{16.28}\right\}\times 100\% = 43\%$$

of the gap.

Thus if the background is social economic status (SES), as Dr. Armor seems to think, then SES explains the race gap almost completely, and there is very little room left for the influence of school-based variables.

3. Criticism

Anybody with any formal training in statistics at all can point out the obvious flaws in this argument. There is generally nothing wrong with the computations, but the interpretation of the results, and the discussion based on these interpretations has nothing to do with either statistics or science. Let us discuss some of the key points in detail.

3.1. **Explaining.** Regression analysis by definition does not "explain" anything. Scientific explanation requires embedding empirical results in a theoretical framework, which can then subsequently be tested empirically. Regression analysis merely establishes correlations between certain selected variables in certain selected groups of students.

3.2. **Causal Attribution.** It is *impossible*, on the basis of a regression analysis with correlated predictors, to state that a certain percentage of the variance of the outcome can be attributed to the effect of predictor A, and another percentage to the effect of predictor B. There is no consistent and logical way to do this.

More generally, analysis such as the one reported by Dr. Armor (and the same thing applies to the other reports as well) are tremendous missed opportunities. The data provided by the district have a fairly complete longitudinal history of the school career of all students in the district, with grades, courses, discipline and so on. Very rich longitudinal analysis of these *event-histories* are possible, and models can be fitted which test causal assumptions.

3.3. **Labeling.** Dr. Armor calls *lunch* a proxy for SES, in fact he seems to treat the early tests in the same way. Since nobody owns exclusive rights to the "correct" definition of SES, there is nothing inherently wrong with this. But it obviously is misleading. In subsequent discussion Dr. Armor argues that he could make the gap even smaller by including more SES proxies. Of course this is true, but it has nothing to do with the "fact" that these are SES variables. They merely have to be correlated with race.

We also emphasize that *lunch* could alternatively be used as a proxy for race. As we shall see in our data analysis below, the correlation between *lunch* and *race* is impressive. It would be equally misleading to speak about *lunch* from now on as if it was really *race*, although this particular labeling makes it very plausible indeed that the race gap almost disappears if we correct it with a proxy for race.

This "naming fallacy" is especially serious, because Dr. Armor includes early test results as an indicator for family background (and thus SES). Again, we might as well use them as a proxy for race. Using test results to correct test results is bound to reduce the gap considerably, because taking standardized tests is a specific skill, both in the first grade and in the eight grade. Thus test results in different grades tend to be highly correlated.

3.4. **Prejudice.** When the gap analysis fails to make the corrected gap statistically nonsignificant, Dr. Armor argues that in his "professional opinion" this is due to unmeasured family influences which continue to operate during the school career. There are, of course, no data presented to support this. Dr.

Armor is so convinced that the remaining gap cannot be caused by anything the school does that he does not need data, only his "professional opinion".

3.5. **Weird Gaps.** There is nothing in the gap analysis procedure proposed by Dr. Armor that guarantees that β_0 is larger than β_1. In fact, it is easy to construct examples, where β_1 is smaller. In such cases, the background "explains" a negative percentage of the gap, whatever that means. Although such examples are perhaps somewhat artificial, they do show that the procedure is fundamentally unsound.

3.6. **Correcting Everything.** In removing the "effect" of the background from the outcome (test) in a regression analysis, we remove at the same time the "effect" of the background on the variable *race*. Thus the gap analysis does **not** quantify the effect of race on test score corrected for background, but it quantifies the effect of *race corrected for background* on *test score corrected for background*. If the background is correlated strongly with race, then the variable *race corrected for background* will be very different from the original race variable, in fact it will have very little interpretable variance left.

3.7. **The Other Gap.** We can follow Dr. Armor and compute the race gap corrected for SES, but we can also proceed the other way around. Compute the raw SES gap (or lunch gap) and correct this for race in exactly the same way. We can associate an equally plausible, or equally silly, story with this analysis. But observe there are some subtle differences here, the main one being that race (i.e. skin color) exists and can be measured unambiguously, while SES is just a theoretical construct. Anybody, and his brother, can design additional proxies for SES and can claim that these variables measure SES in some way or another. As long as there is no objective measurement procedure, we can maintain, with Dr. Armor, that *Free Lunch* is a proxy for SES, and that early test scores (and in other Armor reports even gender) are proxies too. Race is just a single unambiguous variables, taking only two values, while SES is everything the enterprising social scientist says it is.

3.8. **Path Analysis.** It makes more sense to perform a path analysis, or causal analysis, using the diagram below.

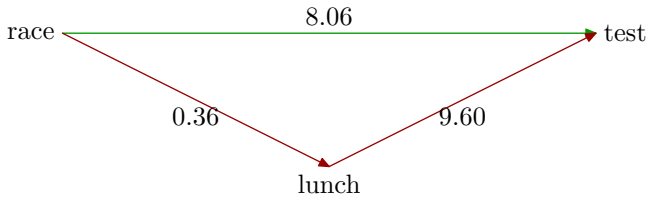


Figure 1. Path Diagram

The diagram tells us that race influences test results in two different ways. The green arrow is the *direct effect* of race, this is identical to the corrected race gap (i.e. the regression coefficient of *race* in a regression of *test* on *race* and *lunch*). But there is also an *indirect effect* of race on test scores, via the free lunch program. This is the product of the two red arrows, which are the regression coefficients of *race* in a regression of *lunch* on *race* and the regression coefficient of *lunch* in a regression of *test* on *race* and *lunch*. For math in the eighth grade the direct effect of race on test is 8.06, the indirect effect is $.36 \times 9.60 = 3.46$, which means that the total effect of race on test scores is 11.52. This total effect approximates the raw gap.

This path analysis is only presented here to illustrate how a more sophisticated gap analysis could be set up. It is in final form, because it assumes, for instance, linear regression of *lunch* on *race*, which is clearly inappropriate. But it does show the pervasive influence of race more clearly than the simple gap analysis does. In fact, the diagram indicates that the effect of race on test score should be estimated using the raw gap (which is the sum of the direct and indirect effects), while the effect of SES on test score should always be corrected for race (because there only is a direct effect).

That this causal model makes some sense is obvious from the fact that we cannot interchange *race* and *lunch* in the diagram. It is nonsense to suppose that being in the lunch program or not determines (is a cause of) someone's race.

4. A Leisurely Look at the Gap

Now let us explain gap analysis somewhat differently. We look at a simple case, and give a simple explanation. We compare the average test scores of four groups:

BL : Blacks in the Free Lunch Program;
BN : Blacks not in the Free Lunch Program;
WL : Whites in the Free Lunch Program;
WN : Whites nit in the Free Lunch Program.

We suppose the average scores of any of these four groups are determined in the following way. All groups have a baseline score μ. The white groups WL] and [WN] get a bonus α, those not in the free lunch program [WN] and [BN] get a bonus β, and whites not in the free lunch program [WN] get an additional bonus γ. Thus we have the following table.

Table 1. Theoretical Expected Values

	black	white
lunch	μ	$\mu+\alpha$
no lunch	$\mu+\beta$	$\mu+\alpha+\beta+\gamma$

The corrected race gap is α, the corrected lunch gap is β, and there is an *interaction* γ. In Dr. Armor's gap analysis, we assume that γ is zero, i.e. there is no interaction.

It follows from the table above that we can estimate

$$\begin{aligned}
\mu &= \text{baseline} = [BL],\\
\alpha &= \text{race gap} = [WL] - [BL],\\
\beta &= \text{lunch gap} = [BN] - [BL],\\
\gamma &= \text{interaction} = [WN] - [BN] - [WL] + [BL].
\end{aligned}$$

If γ is zero, then the table becomes

Table 2. Theoretical Expected Values

	black	white
lunch	μ	$\mu+\alpha$
no lunch	$\mu+\beta$	$\mu+\alpha+\beta$

and we have

$$\begin{aligned}
\mu &= \text{baseline} = [BL],\\
\alpha &= \text{race gap} = [WL] - [BL] = [WN] - [BN],\\
\beta &= \text{lunch gap} = [BN] - [BL] = [WN] - [WL].
\end{aligned}$$

Thus we see that if there is no interaction, then the corrected gap can be computed in two different ways. We can use both $[WL]-[BL]$ and $[WN]-[BN]$ to computed the corrected race gap, the model of gap analysis (without interaction) says that these two estimates should be essentially equal. In Dr. Armor's implementation of gap analysis, the corrected race gap is estimated by using ordinary least squares, which basically takes a weighted average of the two estimates computed by taking differences within rows.

5. Data Analysis

Let us look at a simple situation. Data are selected from the 98-99 data file which contains the 4000 students in the Woodland Hills schools. For each of the students we use the math test *mtn*, the reading test *rtn*, the race

(black, white, or other), the grade, and if they are enrolled in the free lunch program or not. For each grade we look at the lunch $\times$ race table.

This table has two rows (lunch or not) and two columns (black or white – we eliminated the small number of "other" students). Actually, for each grade there are four lunch $\times$ race tables. Both for Math and Reading there is the two-by-two table with frequencies and the two-by-two table of averages. Since

5.1. **Math in Grade 8.** Let us show the actual tables for grade 8. Math first.

Table 3. Cross Table (left) and Math Averages (right)

	black	white
lunch	122	63
no lunch	45	158

	black	white
lunch	33.21	43.22
no lunch	42.44	51.66

We see that of the black students more than 70% is in the free lunch program, while of the white students this less than 30%. The raw race gap is 13.56, the raw lunch gap is 12.97. The corrected race gap is $[WL]-[BL] = 10.01$, while the corrected lunch gap is $[BN] - [BL] = 9.23$. We see that race corrects lunch approximately to the same extent that lunch corrects race. If we estimate the corrected race gaps with regression analysis we find 8.06 for race and 9.60 for lunch, slightly different numbers because of the different estimation method.

Actually, there is yet another table that is of interest. We can also give a two-by-two table of the standard deviations of the four groups in the table.

Table 4. Cross Table (left) and Math StDevs (right)

	black	white
lunch	122	63
no lunch	45	158

	black	white
lunch	16.12	18.02
no lunch	18.17	19.06

For the total group in grade 8 blacks have a standard deviation of 17.00, whites have 19.37. Those in the free lunch program have 17.49, those not in the program have 19.73. Thus it seems that generally there is more variation for whites and more variation for those not in the free lunch program. This systematic difference in the variances of the different cells does violate one of the key assumptions of ordinary linear regression analysis, in particular it invalidates significance tests. This is easy to repair using weighted least squares, but both the fact that there appears to be an interaction in some cases and that the homogeneity of the variances is not true make simple least squares as used by Dr. Armor somewhat suspect.

5.2. **Reading in Grade 8.** The data for reading in eight grade are:

Table 5. Cross Table (left) and Reading Averages (right)

	black	white
lunch	127	64
no lunch	47	159

	black	white
lunch	33.35	43.95
no lunch	41.09	54.85

Observe that the cross table is slightly different from that for math, because students may not take the same tests. We see the corrected race gap for reading is 10.60 (raw gap 16.28), and the corrected lunch gap is 7.74 (raw gap 14.71).

5.3. **Gaps in Other Grades.** Here we give an overview of the raw and corrected gaps for grades 1-8 (still in the 98-99 school year).

Table 6. Math Gaps by Grade

Grade	Lunch Raw	Lunch Corrected	Race Raw	Race Corrected
1	13.70	2.13	18.03	13.58
2	15.62	7.08	16.00	10.71
3	17.00	6.15	13.41	6.55
4	20.99	15.72	18.71	12.45
5	15.86	3.34	18.43	11.28
6	12.91	5.54	16.51	11.84
7	14.17	3.62	19.15	13.91
8	12.97	9.23	13.56	10.01

Table 7. Reading Gaps by Grade

Grade	Lunch Raw	Lunch Corrected	Race Raw	Race Corrected
1	11.16	2.76	14.50	11.56
2	15.69	3.89	15.33	7.89
3	18.21	7.12	12.90	4.97
4	19.26	13.31	18.55	12.82
5	18.54	10.29	20.21	14.62
6	13.84	8.72	12.14	6.76
7	13.98	3.58	16.15	9.60
8	14.71	7.74	16.28	10.60

We see from these tables that in most grades the raw gaps are comparable in size. But in many cases the lunch gap corrected for race is smaller than

the race gap corrected for lunch. Using Dr. Armor's reasoning shows that the effect of SES corrected for race is usually smaller than the effect of race corrected for SES.

We end the analysis by giving plots of the raw and corrected gaps against grade. There is some tendency for gaps (both lunch and race) to be largest around grades 4 and 5, but the effect is not clear. It is very clear, however, that the corrected gap curve follows the raw gap curve very closely, indicating a fairly constant correction percentage.

Figure 2. Raw (green) and Corrected (red) Lunch Gaps for Math

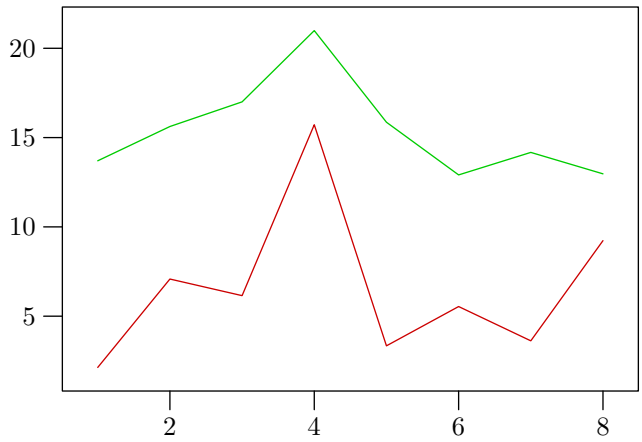


Figure 3. Raw (green) and Corrected (red) Race Gaps for Math

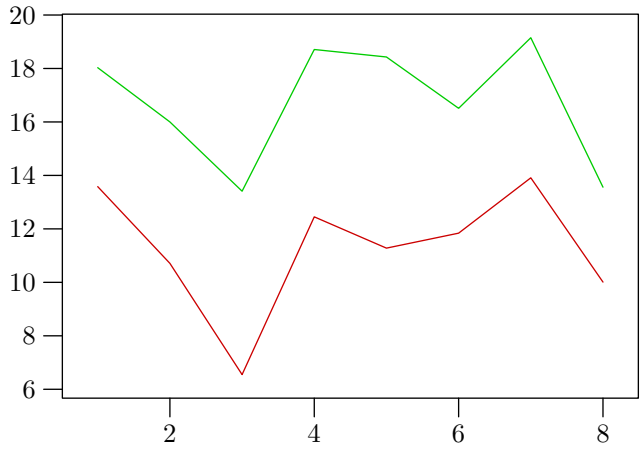


Figure 4. Raw (green) and Corrected (red) Lunch Gaps for Reading

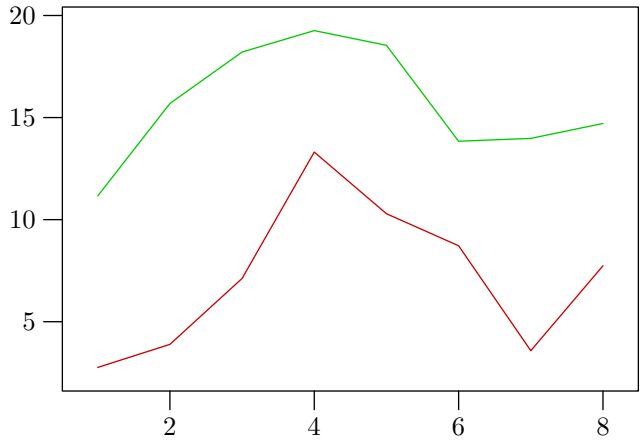
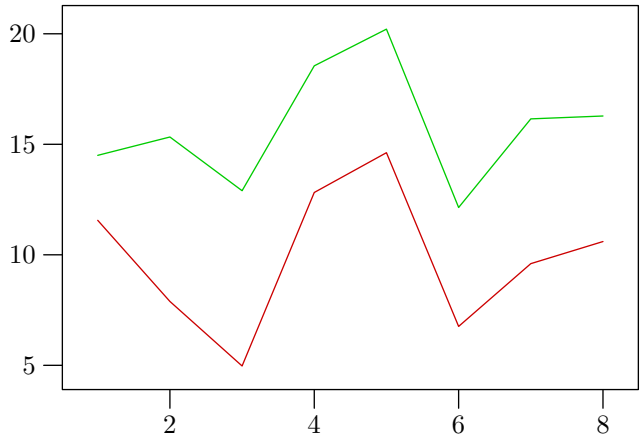


Figure 5. Raw (green) and Corrected (red) Race Gaps for Reading



6. Conclusions

Our main findings from the analysis we have done are the following.

(1) *lunch* reduces the race gap by 40% for mathematics and by 50% for reading. Conversely, *race* reduces the lunch gap by as much as 60% for mathematics and by around 50% for reading.
(2) From the tables there is substantial evidence for interaction between *race* and *lunch*.
(3) The race gap remains substantial after correction for *lunch*.
(4) The SES indicators used by Dr. Armor are so highly correlated with race that they simply cannot be separated, either conceptually or statistically, from race.

(5) There is substantial evidence that black students and students in the lunch program have smaller variability in their test scores.

Appendix A. Remarks on Achilles and Rossell

A.1. **Achilles.** In the report by Dr. Achilles various proportionality factor are computed to compare disciplinary actions in various groups. The big problem with the approach in this paper is that it is essentially univariate or bivariate. A number of factor (race, age, gender, free lunch) are successively related to discipline, but clearly these factors themselves are highly correlated. Thus it is possible that roughly the same conclusion is presented in many different forms. Also, possible interactions are ignored.

The appropriate analysis for data of this kind is a logistic regression analysis, where we predict, on the individual level, if a disciplinary action will be taken. We use race, age, gender, and free lunch as predictors, and we observe the percentage of concordance (how many of our predictions were correct). This will give a much more in depth view of the data, which also does not drown the reader in a flood of unreadable tables. Moreover, the tables invite looking at individual cell entries, i.e. what we normally call data dredging. This is also true, because Dr. Achilles does not do any real statistics, i.e. he does not present standard errors, confidence intervals, or hypothesis tests.

The district database has enough information to do such a logistic regression analysis, but the information necessary to create a database and perform such an analysis was not provided to us in time for this report. The data we received that were used in Dr. Achilles report were only the same tables that are printed in the report. We were not given the underlying data by the Commonwealth.

For an example of what can be done, we refer to Jan de Leeuw, "Regression Analysis in the Wilmington Case", UCLA Statistics Preprint 175, 1994.

A.2. **Rossell.** It was impossible, from the material provided to us, to repeat the analyses done by Dr. Rossell. Moreover, what is really interesting in the enrollment data, are the classes taken by the various groups of students. The Rossell data provided by the Commonwealth did not include classroom enrollment data. Again, there was no usable information in the data provided to us that makes it possible to investigate this. The data are in the districts data base, but they have not been provided to us in a form that can be incorporated in a usable database. They are also provided to us as printed output, but unfortunately printed output cannot be used directly in a statistical analysis.

As usual, the formulas in Dr. Rossell's report are typed in a way that makes it hard to understand them. The dissimilarity index and the interracial

exposure index are not unreasonable, but they aggregate information which would be more useful in tabular form, and they are not really explained. They are "commonly used by social scientists" and they have "been introduced into every school desegregation court case", but it is not explained why they are so ubiquitous.

Jan de Leeuw, Professor and Chair, UCLA Department of Statistics
E-mail address: deleeuw@stat.ucla.edu