A Linear Regression Approach for Simultaneously Adjusting Attributes

Xiao-Li Meng
Department of Statistics, Harvard University

April 12, 2015

Abstract

This note attempts to provide a statistically more principled framework to address the issues raised in the following paragraph:

"Systematic errors in comparing effective areas: Speaking hypothetically, if we label the instruments by numbers $i = 1, \dots, N$ and each has an attribute A that is used to measure the same $j = 1, \dots, M$ astrophysical sources, with intrinsic attribute F_j where $C_{ij} = A_i F_j$ are the instrumental measurements, then the question is: "Is there a way to decide how (or whether) to change A_i when the values C_{ij}/A_i do not agree with F_j to within their statistical uncertainties s_i. In other words, each instrument provides an estimator f_j of F_j with statistical uncertainty s_j but $|f_j - F_j|/s_j$ is often large, not distributed as a Gaussian with unit variance (but can have zero mean if we define $F_j = \sum_j f_j s_j^{-2} / \sum_j s_j^{-2}$). How to estimate the systematic error on the A_i?"

Our key result is that if it is reasonable to assume multiplicative errors (i.e., in terms of percentage) on some initial measurements/estimations a_i for A_i and f_j for F_j, then after observing c_{ij} (as a realization of C_{ij}), we should update/adjust a_i and f_j respectively via "power shrinkage":

$$\hat{A}_i = a_i^{w_b} \times (\tilde{c}_{i\cdot}/\hat{F})^{(1-w_b)}, \quad i = 1, \dots, N \tag{1}$$

$$\hat{F}_j = f_j^{w_g} \times (\tilde{c}_{\cdot j}/\hat{A})^{(1-w_g)}, \quad j = 1, \dots, M. \tag{2}$$

Here $\tilde{c}_{i\cdot}$ and $\tilde{c}_{\cdot j}$ are respectively the ***geometric*** **mean of $\{c_{ij}, j = 1, \dots, M\}$ and $\{c_{ij}, i = 1, \dots, N\}$, w_b and w_g are respectively the relative precisions of $b_i = \log a_i$ and $g_j = \log f_j$, that is,**

$$w_b = \tau_b^{-2}/(\tau_b^{-2} + M\sigma^{-2}), \quad w_g = \tau_g^{-2}/(\tau_g^{-2} + N\sigma^{-2}), \tag{3}$$

where τ_b^2 and τ_g^2 are respectively the variance of b_i and g_j, and σ^2 is the variance of $y_{ij} \equiv \log c_{ij}$, and $\hat{F}$ and $\hat{A}$ are determined by the "self-consistency":

$$\hat{A} = \tilde{a}^{w_b} \times (\tilde{c}/\hat{F})^{(1-w_b)}, \quad \hat{F} = \tilde{f}^{w_g} \times (\tilde{c}/\hat{A})^{(1-w_g)}, \tag{4}$$

where $\tilde{a}$, $\tilde{f}$, and $\tilde{c}$ are respectively the geometric means of a_i's, f_j's, and c_{ij}'s. Further more, the adjustment (1)-(2) are the maximum likelihood estimator when b_i, g_j, and y_{ij} are Gaussian.

1 Potential issues that led to the difficulty ...

There seem to be two issues here. First, there seems to be a mix of *estimators* (calculated from data) and *estimands* (of interest), a mix that is known to cause many problems, e.g., subtracting background noise at the observation level (estimator) can lead to negative counts, whereas subtracting background at the expectation level (estimand) cannot. Second, the estimation approach adopted via computing ratio C_{ij}/A_i is known to be very unstable, and there is a good amount of information in the data that has been ignored because the estimation seems to be carried out on the individual base, instead of simultaneously.

Below I will try to recast this problem as a linear regression, and then fit it via maximum likelihood approach, which is known to produce the most efficient estimator (asymptotically) when the model assumptions hold. The assumptions I made may not be reasonable, so they should be checked, but I expect the results to be more stable than the ratio estimators, even if these assumptions turn out to be unreasonable. Obviously, my recast is based on my very limited understanding of the actual physical process, and hence it may need to be revised seriously.

2 Recasting as a linear regression problem

To avoid the confusion between what can be calculated from data (e.g., estimators) and what are to be estimated (i.e., estimand), we will use lower cases to represent the former, and upper cases for the latter. Hence we will rewrite C_{ij} as c_{ij}, which represents the actual measurement obtained by instrument i from physical source j, and reserve C_{ij} to represent the corresponding but unobserved *true* value if there is no measurement error. Similarly, we will use A_i for the true attribute that makes $C_{ij} = A_iF_j$ holds exactly, but a_i to represent an estimator of A_i. This will be consistent with the notation that f_i is an estimator of F_i.

The key here is to notice that although $C_{ij} = A_iF_j$ holds by our definition, there is no reason to expect that $c_{ij} = a_if_j$, or that c_{ij}/a_i is a good estimate of F_j. Indeed, the ratio estimator c_{ij}/a_i tends to have very large variance because the chance for a_i to be close to zero is much larger than having *both* c_{ij} and a_i close to zero (which is needed to control the variance of the ratio c_{ij}/a_i).

To avoid this problem, we start by noting a trivial fact that $C_{ij} = A_iF_j$ is mathamtically equivalent to

$$\log C_{ij} = \log A_i + \log F_j. \tag{5}$$

However, this relationship holds at the *estimand* level, not at the *estimator/observation* level. Indeed, if we let $y_{ij} = \log c_{ij}, b_i = \log a_i, g_j = \log f_j$, then it is *not* even reasonable to put down the regression model as $y_{ij} = b_i + g_j + \epsilon_{ij}$, and to assume ϵ_{ij} has mean zero and is independent of $\{b_i, g_j\}$. This is because that these assumptions would imply—incorrectly– that the mean of y_{ij} is determined by b_i and g_j, forgetting that they themselves are respectively (bad) estimates of $B_i = \log A_i$ and $G_j = \log F_j$, which determines the mean of y_{ij}.

A statistically more sound way to proceed is as follows. Provided that it is reasonable to assume that the (measurement) errors in x for X, where x can be c, a, or f, are multiplicative (i.e., in terms of percentage), we may postulate the following *three* regression models, where $i = 1, \dots, N$ and $j = 1, \dots, M$.

$$\begin{aligned} y_{ij} &= B_i + G_j \quad + e_{ij}; & (6)\\ b_i &= B_i \qquad\quad + \epsilon_i; & (7)\\ g_i &= \qquad\; G_j \quad + \delta_j. & (8) \end{aligned}$$

Here it is reasonable to assume all error terms $\{e_{ij}, \epsilon_i, \delta_j\}$ have mean zero and are independent of each other, but they may not have the same variances. However, if we believe (5) is a good model (with the correctly specified A_i and F_j), then it will be reasonable to assume that all e_{ij} have the same variance σ^2, that is, $\mathrm{Var}(y_{ij}) = \sigma^2$ for all pairs of (i, j). For ϵ_i and δ_j, we can permit them to have instrumental or source specific variance, as long as these variances can be considered known. But for notation simplicity, we will assume $\mathrm{Var}(\epsilon) = \tau_b^2$ and $\mathrm{Var}(\delta) = \tau_g^2$; this assumption also makes it possible/easier to estimate the variances from the data. But we will proceed by first assuming all the variances as known, and then discuss extended solution when the variance themselves need to be estimated from the data.

3 Fitting the Linear Regression

The regression model given by (6)-(8) is a special case of multivariate linear regression with a particular design matrix, and hence it can be fitted as such. However, it is more straightforward, and instructive, to fit it via maximum likelihood estimation (MLE) method by assuming all errors are independent Gaussians, because it is easy to write down the log likelihood function and then maximize it by setting its relevant partial derivatives (i.e., the score function) to zero. We can then also obtain the variance of our estimators by calculating the Fisher information via taking the corresponding second-order partial derivatives. Derivations will be given in the Appendix. Note the estimators given below will be valid (e.g., consistent, but not necessarily efficient) even when the Gaussian assumption is unreasonable, though in such cases the variance of the estimator requires a more complicated "sandwich" formula, which involves both the Fisher information and the variance of the score function.

To express the MLE in an intuitive way, we first note that for each B_i, there are two pieces of information in the data for estimating it. The direct information comes from b_i, which has precision τ_b^{-2}, the reciprocal of the variance. The indirect information comes from (6), because e_{ij} has mean zero, we would expect that $\bar{y}_{i\cdot} - \bar{G}$, where the average is taken over $j = 1, \dots, M$, would be a good estimator of B_i with precision $M\sigma^{-2}$, *if* $\bar{G} = \sum_{j=1}^{M} G_j/M$ *is known.* The MLE formalizes this intuition by weighting these two estimates proportional to their precisions, and it resolves the issue of unknown $\bar{G}$ via simultaneously estimating all B_i's and G_j's while keeping all the intuitive expressions.

Specifically, let $w_b = \tau_b^{-2}/(\tau_b^{-2}+M\sigma^{-2})$, which is the percentage of precision in the direct information relative to the total precision available for estimating B_i; and similarly $w_g = \tau_g^{-2}/(\tau_g^{-2} + N\sigma^{-2})$, the percentage of precision in the direct information for estimating G_j. Then the MLE for B_i and G_j are given respectively by

$$\begin{aligned} \hat{B}_i &= w_b b_i + (1-w_b)(\bar{y}_{i\cdot} - \hat{G}), \qquad i = 1, \ldots, N, & (9)\\ \hat{G}_j &= w_g g_j + (1-w_g)(\bar{y}_{\cdot j} - \hat{B}), \qquad j = 1, \ldots, M; & (10)\end{aligned}$$

where

$$\begin{aligned} \hat{B} &= \frac{w_b\bar{b} + (1-w_b)w_g(\bar{y}-\bar{g})}{w_b + (1-w_b)w_g}, & (11)\\ \hat{G} &= \frac{w_g\bar{g} + (1-w_g)w_b(\bar{y}-\bar{b})}{w_g + (1-w_g)w_b}, & (12)\end{aligned}$$

and $\bar{b}$, $\bar{g}$, and $\bar{y}$ are respectively the averages of b_i's g_j's and y_{ij}'s.

Consequently., since $b_i = \log a_i$, and $\bar{y}_{i\cdot} = \log \tilde{c}_{i\cdot}$ we see the adjustment from a_i to $\hat{A}_i = e^{\hat{B}_i}$ is given by

$$\hat{A}_i = a_i^{w_b} \times (\tilde{c}_{i\cdot}/\hat{F})^{(1-w_b)}, \quad i = 1, \ldots, N, \tag{13}$$

where $\hat{F} = e^{\hat{G}}$. That is, when we make the adjustment, we will *shrink* the power of the original a_i from 1 to $w_i < 1$, as well as multiply by a factor that is larger than one when $\tilde{c}_{i\cdot} > \hat{F}$ and smaller than one when $\tilde{c}_{i\cdot} < \hat{F}$. Similarly, we adjust f_j to $\hat{F}_j$ via

$$\hat{F}_j = f_j^{w_g} \times (\tilde{c}_{\cdot j}/\hat{A})^{(1-w_g)}, \quad j = 1, \ldots, M, \tag{14}$$

where $\hat{A} = e^{\hat{B}}$. It is worth to compare $\tilde{c}_{\cdot j}/\hat{A}$ with the original c_{ij}/a_i, the former not only use more information in the data, it is much more stable since the $\hat{A}$ is virtually guaranteed to be away from zero. Note it is easy to verify that (11)-(12) are equivalent to (4).

We note that if the variances are unknown, then they can be estimated from the data as well by simultaneously solving (10)-(12), which involve $\hat{\sigma}^2$, $\hat{\tau}_b^2$, $\hat{\tau}_g^2$ because they are needed for computing w_b and w_g, and the following three equations, which involve $\hat{B}_i$ and $\hat{G}_j$:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \hat{B}_i - \hat{G}_j)^2 & (15)\\ \hat{\tau}_b^2 &= \frac{1}{M}\sum_{j=1}^{M}(b_i - \hat{B}_i)^2, & (16)\\ \hat{\tau}_g^2 &= \frac{1}{N}\sum_{i=1}^{N}(g_j - \hat{G}_j)^2, & (17)\end{aligned}$$

The solution $\{\hat{B}_i, \hat{G}_j\}$ and $\{\hat{\sigma}^2, \tau_b^2, \tau_g^2\}$ are then our MLE. If there are any missing values, we can use the EM algorithms, as long as the reasons of missing are known.