Statistical thinking in the simulation design: a continuing conversation on the balancing intercept problem

Boyi Guo, MS, PhD, Linzi Li, MSPH, Jacqueline Rudolph, PhD

Epidemiologists have a growing interest in employing computational approaches to solve analytic problems, with simulation being arguably the most accessible of these approaches. Previous papers have argued the importance of simulation in epidemiology education and research [1, 2]. While these papers discussed the utility of simulation and demonstrated how to carry out simple simulations, few have focused on the statistical concepts underlying simulation approaches. Here, we seek to explicitly connect simulation methods to commonly used statistical concepts, including variable enumeration, generalized linear model, and link functions. Building off the recent series of discussions, we use the problem of the balancing intercept as our motivating example [3]. Specifically, we describe the growing complexity when generalizing to a wider class of data generating mechanisms (e.g., multinomial exposures and covariate adjustment) while demonstrating.

REVIEWING THE BALANCING INTERCEPT

The balancing intercept was first introduced by Rudolph et al. (2021). [3] The goal of the balancing intercept was to set the marginal mean of a simulated variable, using only a simple transformation of the marginal mean as the intercept in the regression model. To explain with a toy example, suppose we are interested in simulating a normally distributed outcome () for two samples (with known group sizes) defined by a binary exposure (). Our goal is to parameterize the simulation based on information in the published literature. In the scenario where the exposure-specific means () are known, we can simply sample the outcomes for each group using these group means and augment the simulated data to form the overall dataset. On the other hand, it is often the case that we can only draw the marginal mean () and the mean difference () between the two groups from published results. Nevertheless, acknowledging the degree of freedom is fixed, we can use and to calculate the group means. This calculation is referred to as the balancing intercept.

|  |  |
| --- | --- |
|  | (1) |

Most (if not all) simulation designs are more complex than a two-sample normal outcome design. For example, we often consider various outcome types (e.g., categorical, survival, or other, complex continuous distributions), multinomial exposures, and estimands of interest. Mirroring real-world observational data, we also routinely need to adjust for cofounders. Equation (1) does not generalize to these complex designs, as was first noticed by Robertson et al [4]. In the following paragraphs, we provide the statistical rationales deciphering these design complications and derive a generalized solution for log-linear models.

CONNECTING SIMULATION TO REGRESSION FUNDAMENTALS

More complex simulation design often requires specifying a data generating model carefully. That data generating model can be rewritten as a series of non-parametric structural equations [2]. For practicality, we often say those functions follow the form of a parametric model, where we set the parameters rather than estimate them from the data. Because we use parametric models, simulation becomes linked to parametric model fundamentals and the myriad options that exist for these models.

Firstly, we discuss why generalizing the type (i.e., distribution) of the outcome and the estimand of interest complicates the calculation of the balancing intercept. As in many empirical studies, it can be reasonable in simulation to specify a generalized linear model to generate particular outcome types. When using such a model form, a link function, , is required to describe the expected mathematical relationship between the exposure and the mean of the outcome in the simulation design. The choice of the link function is highly relevant to the type of outcome and dictated by the estimand of interest. For example, we can simulate a Gaussian outcome with a log function (link function) to study the mean ratio (estimand) or a binary outcome with a logit function to study the odds ratio. When the link function is nonlinear, e.g., log function, the equality between the expectation of a link function and the link function of an expectation no longer holds, . Consequently, the mathematical manipulation in Equation 1 does not apply when the link function is nonlinear. See the [Supporting Information](https://github.com/boyiguo1/Manuscript-Balance_Intercept/blob/master/Manuscript/appendix.pdf) for the complete mathematical reasoning.

Given the appeal of having a simple solution to the balancing intercept problem, is it possible to derive a closed-form equation for non-linear link functions? Yes, but only for a limited number of link functions. For example, in the Supporting Information, we show the derivation of the balancing intercept for the log link function,

.

In contrast, the logit function, another popular link function in simulation, does not have a tractable solution. Hence, we recommend the use of numeric approximation approaches, as proposed by Robertson et al., and Zivich and Ross (2022) [4, 5].

Secondly, we look at how to calculate the balancing intercept when covariates are included in the simulation design. These covariates could be confounders that affect both the exposure and outcome or mediators and effect measure modifiers that affect the outcome. Additionally, we have to consider the complication of including covariates in the simulation design alongside the choice of link function. Building off the equation above, the closed-form equation for log link function when including covariates () generalizes to

.

If the variables, including the exposure, are pairwise independent, the closed-form equation simplifies to

|  |  |
| --- | --- |
|  | (2) |

We can simplify the calculation by replacing the expectation of the exponential function, , with a moment generating function. For example, the moment generating function for normally distributed with mean and variance is . Unfortunately, it is not uncommon for covariates to not be pairwise independent, and not all distributions have a moment generating function, e.g. Cauchy distribution. In these situations, we can apply the Monte Carlo technique to derive Specifically, one can sample the vector of variables with replacement for a large number of iterations (say 1000) and average the exponential function of the randomly sampled data.

Thirdly, one topic that has not been explicitly discussed in the previous papers on the balancing intercept is how to generalize a binomial independent variable, which could either be an exposure or covariate, to a multinomial variable. The challenge here is the enumeration of the categorical variable. When enumerating a binary variable, we create a data column containing zeros and ones to represent the two levels of a variable (implicitly under the referencing coding scheme). Calculating the mean of this enumeration () is straightforward; hence, Equation 1 is mathematically well-defined. Nevertheless, when the variable is multinomial, the mean is not obvious, and Equation 1 fails. Specifically, when enumerating a multinomial variable with levels, we need to create columns in the data matrix to indicate these levels. Each column marks the membership in a corresponding level using either 1 or 0. We can then treat each column as a binary variable and follow the previous equations. However, this approach treats each column as an independent, binary variable and ignores the grouping structure and correlation among columns derived from the same categorical variable, resulting in an inflated approximation of the balancing intercept.

To circumvent this problem, we recommend using the moment generating function to calculate the mean (of a function). For example, if we have a three-level multinomial exposure with the probabilities () for each level and the coefficients on the log scale (), we can use following moment generating function in Equation 2:

.

In the presence of statistical interactions, one can simply treat the statistical interaction as a special case of a multinomial variable by enlisting all possible combinations.

Fourthly, how to enumerate categorical variables are a very relevant concept. Specifically, if changing the variable enumeration would have an impact on the calculation of the balancing intercept. Put another way, the balancing intercept () is the conditional mean of the reference group when the group membership/binary exposure is enumerated using the reference coding system, . The reference coding system has been the implicit default variable enumeration system in the previous balancing intercept discussions [6][7]. The previous balancing intercept literature defaults to the reference coding scheme without mentioning other systems, for example, another commonly used scheme, effect coding. Instead of using 0s and 1s to indicate membership, effect coding uses 0, 1, and -1, emphasizing the deviation from the grand mean, i.e. the average of level means. Under the effect coding scheme, the balance intercept describes the grand mean. When a study is balanced with respect to the exposure variable, the grand mean coincides with the marginal mean. Hence, the marginal mean can be directly used as the balancing intercept and requires no further calculation. In case of an unbalanced design, one can use weighted effect coding instead [8]. However, this simplification has limited utilities - the complications due to or related to the nonlinear link function persist.

SIMULATION EXAMPLE

To demonstrate the closed-form equation (Equation 2), we conducted a simulation study motivated by Robertson et al. (2021). The simulation followed a log-normal model with two variables that are statistically independent, an exposure () and a covariate (). We assumed was a three-level categorical variable with probabilities (0.5, 0.35, 0.15). We examined different distributions for : a Bernoulli distribution with probability 0.8, a continuous uniform distribution bounded between -1 and 3, a standard normal distribution, and a gamma distribution with shape 1 and rate 1.5. We also examined different magnitudes of covariate coefficient ranging from 1 to 3 with 0.5 increments, while fixing the coefficients for the exposure at (0.2, -0.2). We examined targeted marginal means , ranging from 0.1 to 0.9 with 0.1 increments. For each combination of these parameters, we used Equation 2 to calculate the balancing intercept and simulated a dataset of 10,000 observations. We calculated the deviation of the observed mean from the target mean, referred to as bias. We iterated the process 10,000 times and calculated the averaged bias.

Figure 1 shows that the closed-form equation produced unbiased estimates of in the simulated sample.

We also examined when the outcome follows a binomial distribution instead of normal distribution while keeping the log link function. We aimed to examine how these when the outcome is bounded when using an unbounded link function, e.g. log link function, to simulate a bounded outcome, e.g. probabilities or binary outcome. We observed that it is difficult to control the marginal mean with an analytic solution, particularly when the effect size is large. For example, if we run the previously described simulation study with a binary outcome instead of the normal outcome, the same described process would produce a dataset that has a marginal probability that is lower than the target (See supporting information Figure 1).

CONCLUSIONS

In this commentary, we highlighted how standard simulation approaches rely on the fundamentals of statistics and parametric regression, such as link function, expectation calculation, moment generating functions, and the coding scheme. Throughout, we used the balancing intercept as an example to demonstrate these principles. The underlying objective of the balancing intercept, as well as the flaws in its initial development, provided a platform to exemplify these concepts and improve students’ understanding of how they relate to simulation (and regression more broadly). We described how to extend the balancing intercept for various link functions, the inclusion of covariates, and the generalization to multinomial variables. We also derived a close-form equation to calculate the balancing intercept for simulation designs with the log link function. Simulation studies were conducted to demonstrate that the close-form equation produced unbiased estimates of the marginal mean of the outcome.

When introduced in the statistical training required by most epidemiology programs, concepts like alternate coding schemes and moment generating functions can appear merely theoretical or academic – something to be learned for a test but ultimately never used in practice. However, we would like to remind ourselves of the necessity of fundamental statistics training in epidemiology, even in the new era of computation.

Reference

1. Rudolph, J.E., M.P. Fox, and A.I. Naimi, *Simulation as a Tool for Teaching and Learning Epidemiologic Methods.* Am J Epidemiol, 2021. **190**(5): p. 900-907.

2. Fox, M.P., et al., *Illustrating How to Simulate Data From Directed Acyclic Graphs to Understand Epidemiologic Concepts.* Am J Epidemiol, 2022. **191**(7): p. 1300-1306.

3. Rudolph, J.E., et al., *SIMULATION IN PRACTICE: THE BALANCING INTERCEPT.* American Journal of Epidemiology, 2021. **190**(8): p. 1696-1698.

4. Robertson, S.E., J.A. Steingrimsson, and I.J. Dahabreh, *Using Numerical Methods to Design Simulations: Revisiting the Balancing Intercept.* American Journal of Epidemiology, 2022. **191**(7): p. 1283-1289.

5. Zivich, P.N. and R.K. Ross, *RE: “USING NUMERICAL METHODS TO DESIGN SIMULATIONS: REVISITING THE BALANCING INTERCEPT”.* American Journal of Epidemiology, 2022.

6. Muller, K.E. and B.A. Fetterman, *Regression and ANOVA: An integrated approach using SAS software*. 2002: SAS Publishing. pp. 299-312

7. Lash, T.L., et al., *Modern Epidemiology*. 4 ed. 2021, Philadelphia, PA: Lippincott Williams and Wilkins. pp. 407-408

8. Te Grotenhuis, M., et al., *When size matters: advantages of weighted effect coding in observational studies.* International Journal of Public Health, 2017. **62**(1): p. 163-167.

Figure 1: The derived closed-form equation controls the marginal mean

at the target level for log-normal model

