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 among all 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 them, few have focused on connecting underlying statistical concepts to these simulation approaches, creating gaps between theory and application. Here, we seek to put commonly used statistical concepts, including variable enumeration, generalized linear model, and link functions, in the context of simulation methods. Based on the recent series of discussions on the balancing intercept [3], we explain the growing complexity when generalizing the balancing intercept to a wider class of simulations and revise the closed-form equation for the balancing intercept under assumptions.

# REVIEWING THE BALANCING INTERCEPT

The balancing intercept, first introduced by Rudolph et al. [3], is an estimation of the unknown intercept in regression-based data-generating mechanisms to control the marginal mean of a simulated variable. To explain with a toy example, suppose we are interested in simulating a normally distributed outcome () conditioning on a binary exposure (with known group sizes. Our goal is to parameterize the simulation using the mean difference between the exposure groups (), commonly reported in the literature, such that the marginal mean is fixed at a level of interest. Without the group means directly available, one needs to calculate the intercept (­in a regression system . Acknowledging the degree of freedom is fixed, Rudolph et al. (2021) provided a closed-form equation to calculate the , referred to as the balancing intercept,

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

Obviously, 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), continuous or multinomial exposures (i.e. exposures with more than 2 levels), and different 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].

# CONNECTING SIMULATION TO REGRESSION FUNDAMENTALS

Complex simulation designs require specifying a data-generating model carefully, normally written as a series of structural equations [2]. For simplicity, those equations are assumed to follow some parametric forms and are often expressed as a generalized linear model (GLM). Hence, familiarizing with the concepts of GLM can address challenges when deriving the balancing intercept in more complex data-generating models.

## Outcome and Estimand

Given the type of outcome, data are sampled from a distribution, most commonly Gaussian distribution for continuous outcomes, Bernoulli distribution for binary outcomes, and Weibull distribution for survival time. In addition, a link function, , that describes the expected mathematical relationship between the exposure and the mean of the outcome is specified 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 binary outcome using a logarithmic (log) function to study the risk ratio (estimand), a logit function to study the odds ratio, or a identify function to study the risk difference.

Different choices of the link function lead to different complexities to derive the balancing intercept, with a linear function being the easiest to calculate (see Equation (1)). 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, , as shown in Jensens’ inequality. Consequently, Equation (1) does not hold for non-linear link functions. See the [Supporting Information](https://github.com/boyiguo1/Manuscript-Balance_Intercept/blob/master/Manuscript/appendix.pdf) for the complete mathematical reasoning.

Is it possible to derive a closed-form equation for non-linear link functions? Yes, but only for a limited number of link functions. When applying a log link function, the balancing intercept is calculated (derivations shown in [Supporting Information](https://github.com/boyiguo1/Manuscript-Balance_Intercept/blob/master/Manuscript/appendix.pdf)) via

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

In contrast, the logit function, another popular link function in simulation designs, does not have a tractable solution. Hence, we recommend using numeric approximation approaches, discussed in Robertson et al. [4] and Zivich and Ross [5].

## Covariate

Statistical covariates are one of the most fundamental concepts in data analysis. The covariates could be confounders that affect both the exposure and the outcome, or mediators and effect measure modifiers that affect the outcome. When considered in simulation designs, the multivariate distribution of exposure and covariates needs to be accounted for and increase the complexity of the balancing intercept calculation.

When the variables, both the exposure and the covariates , are pairwise independent, the joint mean can be written as a function of marginal means and For example, when applying to simulation designs with the log link functions, Equation (2) extends as

|  |  |
| --- | --- |
| . | (3) |

We can simplify this calculation by replacing the expectation of the exponential function, , with its moment generating function. For example, the moment generating function for normally distributed with mean and variance is .

Confounding and effect mediation set up hierarchical structures in the simulation of statistical covariates and undermine the pairwise independent assumption. In addition, not all distributions have moment generating functions, 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.

## Multinomial covariates and Coding Schemes

Does having a multinomial exposure, in contrast to a binary exposure, complicate the calculating balancing intercept? The short answer is surprisingly no due to its discrete nature. When enumerating a nominal variable with levels, we normally compose a data matrix (denoted as ) where each column represents one level of this nominal variable. The expectation for the nominal variable with the sampling probabilities () with the corresponding effect coefficients () on the appropriate scale (link function), , can be expressed as . 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.

The enumeration of each level () in this multi-column matrix is called a coding scheme. The most popular coding scheme, as well as the default in the previous balancing intercept literature, is the reference cell coding (also known as dummy coding). With the reference cell coding, the data matrix consists of column to indicate these levels, assuming the reference level collapse with the intercept term. Each column marks the membership in a corresponding level using either 1 or 0. The expectation can be simply written as with . There exist other coding schemes that can possibly simplify the calculation of the balancing intercept. For example, effect coding encodes different levels with a combination of 0, 1, and -1, with the coefficients emphasizing the deviation from the grand mean, i.e. the average of level means. When a study is balanced with respect to the exposure variable, the grand mean coincides with the marginal mean. Hence, the balancing intercept is the targeted marginal mean under the effect coding scheme, requiring no further calculation. In case of an unbalanced design, one can replace effect coding with the weighted effect coding [8].

# SIMULATION EXAMPLE

To demonstrate the closed-form equation (Equation (3)), we conduct a Monte Carlo simulation study motivated by Robertson et al. (2021). The simulation follows 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 the sampling probabilities (0.5, 0.35, 0.15). We examined different distributions for : (1) a Bernoulli distribution with probability 0.8, (2) a continuous uniform distribution bounded between -1 and 3, (3) a standard normal distribution, and (4) a gamma distribution with shape 1 and rate 1.5. We also examine 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 evaluate multiple targeted marginal means , ranging from 0.1 to 0.9 with 0.1 increments, where is sampled from a normal distribution with standard deviation 0.1. For each combination of these parameters, we use Equation (3) to calculate the balancing intercept and simulated a dataset of 10,000 observations. We calculate the deviation of the observed mean from the target mean, referred to as bias. We iterate 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 examine the bias when applying to the binomial outcome , instead of the normal distribution, while keeping the log link function. (See supporting information Figure 1) We observe that it is difficult to control the marginal mean with an analytic solution of balancing intercept, particularly when the effect size is large. This can be explained by the fact that the outcome is bounded, e.g. probabilities or binary outcomes, while the link function is not, creating asymmetry that skews the distribution of the mean.

# CONCLUSIONS

In this commentary, we highlight 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. We describe how to extend the balancing intercept for various link functions, the inclusion of covariates, and the generalization to multinomial variables. We also derive 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, statistical concepts like 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. Under the context of the balancing intercept problem, we show these statistical concepts can address computational challenges in real-world problems. The balancing intercept problem provides a didactic platform to exemplify these concepts and a case study that facilitates students to understand elements of data-generating models. 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

