Lean on your statistics: the simplification of the balance intercept problem

Boyi Guo, Jacqueline Rudolph

2022-05-24

As the field of epidemiology evolves, there are growing interest to employ more computational approaches to solve analytic problems. Among them, simulation is one of the most accessible concept. Previous literature argues the importance of simulation in epidemiology education and research. [TODO: add citations] Even though computational tools can be very helpful, we caution the excess reliance on the computation in analytic problem solving and the neglect of fundamental statistics theory. In the article, we demonstrate how a basic statistics knowledge can simply analytic problems by visiting a specific simulation problem, the balance intercept.

The balance intercept problem was first introduced by Rudolph et al. (2021) to addressed the problem of controlling the marginal probability of binary outcomes when constructing a simulation study. The authors proposed to numerically calculate the "balance intercept" to replace the "standard intercept" in simulation procedures. This same problem was later revisited by Robertson, Steingrimsson, and Dahabreh (2021) who discovered that the analytic solution of the balance intercept analytically can produce inaccurate controlling of the marginal probability at the desired level. Instead, Robertson, Steingrimsson, and Dahabreh (2021) proposed a numeric solution to solve for the balance intercept for binomial simulation with a logistic link function. Later, Zivich and Ross (2022) did xxx for multi-level categorical variable. [TODO: add citence]

To better understand the balance intercept problem and the proposed simplification, we first introduce a basic statistics concept, coding scheme. A coding scheme describes how a categorical variable is enumerated in the regression system. Most commonly used coding schemes in analysis include reference coding (also known as dummy coding) and the effect coding. Both coding schemes create $p-1$ binary columns for a categorical variable with p levels. The reference coding employs 0 and 1 to denote the level an individual belongs, while the effect coding normally employs 0, 1, and -1. As the enumerations are different for the two schemes, they offer different interpretations of regression coefficients. (See example in Table) The reference coding emphasizes the change relative to a reference level of preference; the effect coding emphasizes the deviation from the grand mean (here refer to as the mean of the means). Nevertheless, the two schemes translate to each other one-on-one, and hence provide the same statistical inference. To note, both coding schemes can be applied ubiquitously in any regression system regardless the outcome distribution and the link function of choice. For more techinical details, we defer to [TODO: add textbook, probably INTRODUCING ANOVA and ANCOVA by andrew rutherford].

The balance intercept problem embeds in the reference coding system in the original proposal to control the marginal probability, even though the authors didn't mention explicitly. The intercept term in the reference

Levels	Referece			Effect		
	X_1	X_2	Conditional Probability	X_1	X_2	Conditional Probability
Level 1	0	0	β_0	1	0	$\beta_0+\beta_1$
Level 2	1	0	$\beta_0+\beta_1$	0	1	$\beta_0+\beta_2$
Level 3	0	1	$\beta_0+\beta_2$	-1	-1	$\beta_0-\beta_1-\beta_2$

Table 1: One-on-one translation between the reference and effect coding and their calculation of conditional probability of a three-level categorical variable.

coding scheme describes the conditional probability of the reference level. Hence, Setting the intercept term to the target marginal probability (referred to as the standard intercept) would fail to control the empirical marginal probability due to the theoretical discrepancy. The balance intercept is a derivation of the conditional probability of the reference level, and can be calculated with previous solutions. Nevertheless, the accuracy of the simulation study would be vulnerable to the quality of numeric analysis and programming efficiency.

The process of finding the balance intercept may not be necessary when the effect coding system, and hence greatly reduces the amount of numeric computation. The intercept term in the effect coding scheme describes the mean of the group means, which coincide with the marginal probability when the sample sizes are balanced across groups. In other words, the balance intercept problem does not exist for simulations with balanced designs. When the underlying design is not balanced, some simple arithmetic calculation is still needed to calculate the grand mean. However, this can be done easily with a closed form equation and applied to any data generating model of interest, following the mathematical definition of the intercept terms. Besides its straightforwardness, the proposed solution should not incur any knowledge burden as the effect coding is commonly introduced in introductory statistics classes. We demonstrate the simulation with the following toy example.

Balance Intercept of effect Coding

1. Treatment of continuous variable
2. Treatment of categorical variables
3. Generalizability of the approach

Here, we adapt the toy example in xxx to illustrate the difference and interpretation of the two coding scheme.s

Balanced design

It is very simple to control the marginal probability for balanced design when the simulation is based on the effect coding. No calculations is required as the intercept in an effect coding model is the grand mean, in the binary outcome case, the marginal probability. Here we illustrate the process following the simple additive probability example in Rudolph et al. (2021), where the target marginal probability is 0.3, the effect size as in risk difference is 0.2, and a balanced covariate with 2 levels.

```
set.seed(123)

n <- 10000

# Marginal probabilities of each variable
p.y <- 0.3
p.x <- 0.5
rd <- 0.2

# Example 2: Generate L, X, and Y -------------------
X <- rbinom(n, 1, p.x)

# Generate X with marginal prob 0.5
dev_coding <- contr.sum(2)  # Deviation Coding with 2 levels
X_design_dev <- cbind( 1,   # Adding intercept column
                       dev_coding[X+1,]) # Construct the deisgn matrix
```

```
# The design matrix with effect coding can be more easily construct with model.matrix function

beta_vec <- c(p.x,   # Intercept term, the marginal probability for balanced desgin
              -rd/2) # Set up conditional prob for reference level
Y <- rbinom(n, 1, X_design_dev %*% beta_vec)
```

To validate the simulation, we can see the marginal probability 0.5041, and the conditional probability of the two levels are 0.3985 and 0.6122 respectively. Hence the simulation matches with the desired design.

```
mean(Y);
```

```
## [1] 0.5041
```

```
mean(Y[X==0]);
```

```
## [1] 0.3984576
```

```
mean(Y[X==1]);
```

```
## [1] 0.6121788
```

One can use the following code to examine data quality from the modeling perspective.

```
summary(glm(Y~X, family = binomial(link="identity"))) # Reference coding model
summary(glm(Y~X_design_dev-1, family = binomial(link="identity"))) # Effect coding model
```

Unbalanced design

When the groups are not balanced, the simulation with effect coding is less straightforward compared to the balanced case, mainly because the equality between intercept and grand mean doesn't hold. The intercept needs to be adjusted based on the conditional probability of one of the levels (default to the reference level in the reference coding scheme). This adjustment of intercept requires some arithmetic calculation; nevertheless, in the author's biased view, the complexity is still manageable and requires less calculation than the previous proposals. We demonstrate the simulation procedure for unbalanced design with a toy example. The simulation settings are similar to the example above except that we change the group ratio to 8:2 and the effect size to 0.4.

To calculate the new intercept, we need first to establish the conditional probability for one of the levels (by default the $X = 0$ level in this example). As we know that the marginal probability can be expressed

$$Pr(Y = 1) = \frac{n_1 Pr(Y = 1|X = 0) + n_2 Pr(Y = 1|X = 1)}{n_1 + n_2} = \frac{n_1 Pr(Y = 1|X = 0) + n_2(Pr(Y = 1|X = 0) + RD}{n_1 + n_2}$$

where RD is the effect size in risk difference, n_1 and n_2 are the group sample size for $X = 0$ and $X = 1$ respectively. Given $Pr(Y = 1), n_1, n_2$ and RD, we can derive easily derive the conditional probability of X=0,

$$Pr(Y = 1|X = 0) = \frac{(n_1 + n_2)Pr(Y = 1) - n_2 RD}{n_1 + n_2}.$$

The intercept, a_0, as the mean of the group means can be calculated with

$$a_0 = \frac{2Pr(Y=1|X=0) + RD}{2}.$$

The simulation procedure translates to the toy example as

```
set.seed(123)

n <- 10000

# Marginal probabilities of each variable
p.y <- 0.3
p.x <- 0.8      # Imbalanced design
rd <- 0.2

cond.p <- (n*p.y - n*(p.x)*rd)/n
a.0 <- cond.p + rd/2


# Example 2: Generate L, X, and Y -------------------
X <- rbinom(n, 1, p.x)

# Generate X with marginal prob 0.5
dev_coding <- contr.sum(2)  # Deviation Coding with 2 levels
X_design_dev <- cbind( 1,   # Adding intercept column
                       dev_coding[X+1,]) # Construct the deisgn matrix

# The design matrix with effect coding can be more easily construct with model.matrix function

beta_vec <- c(a.0,   # Intercept term, the calculated mean of group means
                -rd/2) # Set up conditional prob for reference level
eta <- X_design_dev %*% beta_vec
eta[eta<0] <- 0
eta[eta>1] <- 1
Y <- rbinom(n, 1, eta)
```

A quick examination shows the simulated data matches with the expectation.

```
mean(Y);
```

```
## [1] 0.2976
```

```
mean(Y[X==0]);
```

```
## [1] 0.1380195
```

```
mean(Y[X==1]);
```

```
## [1] 0.3362315
```

```
# summary(glm(Y~X, family = binomial(link="identity"))) # Reference coding model
# summary(glm(Y~X_design_dev-1, family = binomial(link="identity"))) # Effect coding model
```

To note, if we ignore the group ratio change and don't adjust the intercept in the simulation procedure (for example, use the balanced-design simulation procedure without any modification), the observed marginal probability deviates from the target marginal probability and the deviation is more obvious with more extreme value of effect size and unbalanced group ratio.

Conclusion

In this report, we propose to use the effect coding scheme in simulations to address the problem of controlling marginal probability of binary outcomes. We provide preliminary evidence that the proposal works for both balanced and unbalanced designs via toy examples. Compared to the previous solutions that based on the reference coding scheme, our proposed solution requires less calculation than the approaches to derive analytic and numeric approximation of the balance point. Particularly, it requires modest calculation when the study design is balanced.

The problem of controlling marginal probability in essence is to find the conditional probability for the reference group, as all the other simulation parameters and the model degrees of freedom are considered fixed.

In this report, we only consider easy simulation scenarios, i.e. binary covariates, one covariate, identify link function, to demonstrate the feasibility of this simulation strategy. We anticipate with the levels of a covariate and the number of covariates growing, the calculation complexity would grow but still be manageable.We will provide a more delicate equation, particularly for the unbalanced design, to generalize for those situation. We will also conduct larger scale of simulation studies to evaluate the efficacy of the proposed solution.

The reason why this is necessary * significantly reduces the complexity of the problem, leveraging a statistics concept covered in the introductory level statistics class * Require less numeric programming, which can easily produce errors for people without numeric programming background. The simulation strategy translates easily to other programming e.g. *SAS*, *STATA*, whose implementations are not shown in this article.

The author doesn't arugula if analytic approach and computational approach is superior. Instead, we advocate for a balanced emphasis on both computational skills and analytic thinking.

An alternative simulation strategy to control the marginal probability of a binary outcome. Instead of constructing the simulation based on the reference coding scheme, we encourage using effect coding, specifically deviation coding, to construct the design matrix of categorical covariates. The theoretical basis of this proposal is that the intercept term of the effect coding model (regardless of the link function or parametric assumptions) is the mean of the group means, which coincides with the marginal probability of a binary outcome when the groups are balanced. Hence, to simulate data from balanced designs, no additional calculation is needed compared to the previous proposals. In the case of unbalanced design, it is very intuitive to adjust the simulation equations and requires minimum arithmetic calculation. (See examples below)

References

Robertson, Sarah E, Jon A Steingrimsson, and Issa J Dahabreh. 2021. "Using Numerical Methods to Design Simulations: Revisiting the Balancing Intercept." *American Journal of Epidemiology*, November, kwab264. https://doi.org/10.1093/aje/kwab264.

Rudolph, Jacqueline E, Jessie K Edwards, Ashley I Naimi, and Daniel J Westreich. 2021. "SIMULATION IN PRACTICE: THE BALANCING INTERCEPT." *American Journal of Epidemiology* 190 (8): 1696–98. https://doi.org/10.1093/aje/kwab039.

Zivich, Paul N, and Rachael K Ross. 2022. "RE: "Using Numerical Methods to Design Simulations: Revisiting the Balancing Intercept"." *American Journal of Epidemiology*, May. https://doi.org/10.1093/aje/kwac083.