Research Article
Volume 14:03, 2023

Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Open Access

Development of Composite Index in Clinical Research

Peijin Wang, Weijia Mai* and Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, North Carolina, USA

Abstract

In clinical research, it is often of interest to build a medical predictive model based on a number of independent variables (predictors) which are considered most relevant to the dependent variable (clinical outcome). Baseline demographics and patient characteristics which may inform disease status and/or treatment effect are often considered relevant predictors. These predictors, however, may be highly correlated. In the interest of parsimony of predictors (or least dominated parameters), a composite index is usually developed which combines highly correlated predictors into a single predictor. The exponential-type composite index is commonly seen in clinical research, for example, the body mass index (BMI). In this article, several statistical methodologies for the development of exponential-type composite index are derived, including the multiplicative model and additive model. The relative performance of these methods are compared via extensive clinical trial simulations. Both multiplicative and additive composite index can successfully inform disease status and/or detect treatment effect. Multiplicative composite index has smaller standard error of estimated coefficient; whereas the additive composite index is more beneficial in making accurate predictions. Both methods have higher empirical statistical power for partial F test than situations without using any composite index

Keywords: Composite index • Multiplicative model • Additive model • Modified Gauss-Newton method

Introduction

Composite index is a single variable that aggregates information from two or more measurements (independent variables) that are highly correlated either statistically or clinically [1]. The composite index is often used in clinical research to inform disease status and/or detect treatment effect. It provides a summary evaluation of complex features and/or deals with statistical inference issues. A typical example of composite index in clinical research is the body mass index (BMI). BMI combines weight and height into a single index which can best inform the obesity (disease status) and treatment effect (reduction of obesity) of the participants under study. Statistically, the idea of combining highly correlated independent variables into a composite index can alleviate multi-collinearity issue in regression model. Multi-collinearity among independent variables (predictors) is a common problem in regression model, especially in the field of clinical research [2]. One of the major reasons for the existence of multi-collinearity is the correlations existing among predictors. In the presence of multi-collinearity, the commonly used ordinary least squares (OLS) methods yield unstable estimates of regression coefficients. That is, coefficient estimations tend to have large standard errors, which contaminate the reliability of statistical inference [3]. Assuming the highly correlated predictors are redundant measures for the same underlying theoretical structure, building a composite index is one possible solution to multi-collinearity [4].

This assumption guarantees the validity of creating a composite index, i.e., the index provides meaningful information [1]. By replacing the highly correlated predictors with the proposed composite index, multi-collinearity can be alleviated. The idea of using a composite index to measure multi-dimensional features is well-accepted and wildly used in clinical research. Under linearity assumption, one of the most common applications of composite index is to generate a “score” to measure disease activity and patient health status by taking the sum of individual predictors [5 -7]. In this case, the scale of each predictor can greatly impact the “score” when individual predictors are continuous. To avoid being influenced by the scale and allowing each predictor to have different weight, Andrade C [2] suggested using the weighted sum of Z scores to establish a composite index.

Now the question becomes how to determine the weight for each variable. Principal component analysis (PCA) is a favourable approach in dimension reduction, especially when there are a large number of independent variables. Scholars have used PCA to propose composite index which contained thousands of variables [8,9]. However, as an unsupervised approach, PCA doesn’t use any information from the outcome variable, which makes it less favorable in predictive models with composite index and composite index using PCA to determine weights works well only when individual predictors are highly correlated, otherwise the statistical power will be contaminant [1]. A better alternative in determining the weights is the partial least squares (PLS) approach. Instead of only maximizing sample variance as PCA, the objective function of PLS is to maximize both correlation to the dependent variable and the sample variance. Composite index proposed by PLS was shown to have very good performance, e.g., it has higher correlation to the outcome than the composite index from PCA [10,11].

Based on the idea of making use of the outcome, the latent variable model was proposed to create composite index for longitudinal data, which was shown to have better performance than the composite score approach in terms of type I error and statistical power [12,13] Another way to make use of the dependent variable in creating composite index is the regression model. By fitting the linear regression model, researchers can use either coefficients of variables or Pearson correlation ratio to determine the weight of each variable [14-16]. Though a natural extension from linear structure composite index is the nonlinear ones, few scholars have investigated methods to propose a nonlinear composite index. Generalizing the Pearson correlation ratio of linear regression models, Becker W, et al [17] proposed to use nonlinear correlation ratio to determine the weights. Specifically, the nonlinear model was fitted via Gaussian process; however, Gaussian process is computationally inefficient. For some simple nonlinear structures, scholars tried to transform a nonlinear model into a linear model to simplify calculation. One commonly seen simple nonlinear structure in clinical research is the exponential structure. For instance, BMI for measurement of obesity is defined as , Weight /Height^2 and the QTc interval for assessment of cardio toxicity can be defined as $QT/\sqrt{RR}$ or $QT/\sqrt[3]{RR}$ [18,19]. This structure sometimes can be more interpretable, such as BMI can be treated as the body mass.

**Address for Correspondence: Weijia Mai, Department of Biostatistics and Bioinformatics, Duke University School of Medicine, North Carolina, USA, Tel: 2066614032, E-mail: weijia.mai@duke.edu*

Received: 25 May, 2023, Manuscript No. jbmbs-23-100097; **Editor assigned:** 26 May, 2023, PreQC No. P-100097; **Reviewed:** 09 June, 2023, QC No. Q-100097; **Revised:** 14 June, 2023, Manuscript No. R-100097; **Published:** 30 June, 2023, DOI: 10.37421/2168-9768.2023.14.155

This article innovatively utilizes exponential regression model to propose an exponential-type composite index with data-driven weight for each variable. Building on an earlier proposal of using regression coefficients of some authors, see Chow et al., the weight of each predictor in the composite index is determined by fitting the nonlinear (exponential) regression model. Two methods of measuring the weights are proposed: the multiplicative model and the additive model. Under multiplicative model, after some algebraic transformation, the weights are determined using OLS; and under additive model, the weights are calculated using the modified Gauss-Newton method. The advantages of the proposed exponential models are (i) the proposed exponential-type composite index coincident with many well-established and well-accepted clinical composite index; (ii) this composite index is more accessible, especially when there are not many highly correlated predictors in the regression model; and (iii) the proposed methods are more computationally efficient and easy to implement than some of the above-mentioned methods. Here, the proposed methods are comprehensively evaluated in terms of the probability of accurately informing disease status, regression predicted values, estimated coefficient standard errors, and statistical power of partial F tests. In Section 2, a statistical procedure to generate composite index and potential difficulties are explained. Section 3 provides the statistical methods used to propose an exponential-type composite index under multiplicative and additive models. Section 4 discusses the evaluation methods for composite index. Section 5 contains the simulation result in comparing the performance of each method. Finally, some concluding remarks are shown in Section 6.

Composite Index

One major application of composite index in clinical research is the regression model with highly correlated predictors. Consider a multivariate linear regression model

$$Y=\beta_0+\beta_1X_1+\ldots+\beta_pX_p+\beta_{p+1}X_{p+1}+\ldots+\beta_mX_m+\varepsilon, \tag{1}$$

where Y is the response variable, $X_{1\,,}$ … , X_m are predictors, and $\beta_0,\beta_1,\ldots,\beta_\mu$ are regression coefficients. One typical assumption of the random error term is that $\varepsilon\sim N(0,\sigma^2)$. This implies that Y is assumed to follow $N(\beta_0+\beta_1X_1+\ldots+\beta_mX_m,\sigma^2)$. Suppose there are p highly correlated independent variables in the regression model, without loss of generality, assume the highly correlated variables are $X_1,\ldots,X_P$ (P‹M). Then the predictors can be categorized into two clusters: the high-correlation cluster{ $X_1,\ldots, X_P$ } and the low-correlation cluster { $X_{p+1},\ldots, X_m$}

The composite index is proposed to replace $X_1,\ldots,X_P$ in original regression model (as Eq.(1)), which can be treated as a function of these variables, i.e., the composite index is *$C=G\ (X_1,\ldots,X_P)$,* and *g* is the utility function representing the relationship of these highly correlated predictors . g can be either linear or nonlinear. In this paper, an is considered, that is $g(X_1,\ldots,X_p)=e^{\gamma_0}X_1^{\gamma_1}\ldots X_p^{\gamma_p}$, where $\gamma_0,\gamma_1,\ldots,\gamma_p$ are coefficients needed to be estimated.

The "weights" (or coefficients) in the utility function can be estimated by fitting the following nonlinear regression model [18] :

$$Z\sim e^{\gamma_0}X_1^{\gamma_1}X_2^{\gamma_2}\ldots X_p^{\gamma_p}, \tag{2}$$

where $X_1,\ldots, X_P$ are independent variables, Z is the dependent variable, and observations for Z are obtained by fitting the original regression model in Eq.(1) using OLS. Let $\beta_0,\beta_1,\ldots,\beta_m$ denote the estimated coefficients, and observations $\tilde{z}$ can be derived as following.

$$\hat{Z}=\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p, \tag{3}$$

and $\tilde{z}$ can be treated as information in the regression model explained by $X_1,\ldots, X_P$

The reason for not using equal sign in Eq.(2) is that, to make it a valid regression model, some additional assumptions on the error term and/or some transformations need to be applied to *Z* to estimate coefficients. Now, suppose the way to determine the "best" estimated values for $\gamma_0,\gamma_1,\ldots,\gamma_p$ is proposed, then the composite index can be derived by plugging in the estimated coefficients as following:

$$C=e^{\gamma_0}X_1^{\gamma_1}X_2^{\gamma_2}\ldots X_p^{\gamma_p} \tag{4}$$

The regression model in Eq.(1) can be rewritten using the proposed composite index:

$$Y=\beta_0+\beta_cC+\beta_{p+1}X_{p+1}+\ldots+\beta_mX_m+\varepsilon. \tag{5}$$

In summary, the composite index can be derived as the following steps:

Fit the multivariate regression model $Y=\beta_0+\beta_1X_1+\ldots+\beta_mX_m+\varepsilon$.

Obtain estimated coefficients for β's using OLS and compute $\tilde{Z}$.

Fit nonlinear regression model $Z\sim e^{\gamma_0}X_1^{\gamma_1}\ldots X_p^{\gamma_p}$, and obtain the estimated coefficients for $\tilde{a}_0,\tilde{a}_1,\ldots,\tilde{a}_p$.

Propose the composite index $C=e^{\gamma_0}X_1^{\gamma_1}X_2^{\gamma_2}\ldots X_p^{\gamma_p}$.

It should be noted that under normality assumption of random error term in Eq.(1), both γ and estimates of follow normal distribution, and so does **Z.** For nonlinear regression model in Eq.(2), since the right-hand side has an exponential-type form, one typical approach is to take logarithm and transform it into a linear model, then use OLS to estimate coefficients. However, after taking logarithm, in **Z** will no longer follow a normal distribution, which makes OLS inapplicable. The next section will discuss how to set up and fit the nonlinear regression model in Eq.(2), and finally derive a valid composite index.

Nonlinear/Exponential models

Graver and Boren [20]. Categorized nonlinear regression models with exponential form $\varepsilon^{\gamma_0}X_1^{\gamma_1}X_2^{\gamma_2}\ldots X_p^{\gamma_p}\delta$ into two different categories: the multiplicative models and the additive models. The primary difference between these two models is how the random error term is considered in the nonlinear model. One major assumption of the exponential form regression model is that all of observations for predictors are positive, and dependent variable Z follows normal distribution. That is, the highly correlated predictors in the original model must be strictly positive.

Multiplicative Model

The multiplicative model assumes that the nonlinear regression model has form $e^{\gamma_0}X_1^{\gamma_1}X_2^{\gamma_2}\ldots X_p^{\gamma_p}\delta$ where δ is the random error term, and δ is assumed to follow log-normal distribution. As mentioned previously, one typical approach to estimate coefficients is to take logarithm, then use OLS. Since ***Z*** follows normal distribution, e^Z follows log-normal distribution. The multiplicative model can be written as

$$e^Z=e^{\gamma_0}X_1^{\gamma_1}X_2^{\gamma_2}\ldots .X_p^{\gamma_p}\delta \tag{6}$$

Then, $\tau=\ln\delta$ follows normal distribution. Denote the distribution for τ as $\tau\sim N\left(\mu,\sigma_M^2\right).$, and the mean and variance for δ are derived as

$$E(\delta)=e^{\mu+\frac{\sigma_M^2}{2}}, \text{ and} Var(\delta)=\left(e^{\sigma_M^2}\text{-}1\right)e^{2\mu+\sigma_M^2}. \tag{7}$$

To have $E\left(e^Z\right)=e^{\gamma_0}X_1^{\gamma_1}X_2^{\gamma_2}\ldots X_p^{\gamma_p}$, $E(\delta)$ must be 1. Then, $\mu=-\sigma_M^2/2$ and $Var(\delta)=e^{\sigma_M^2}\quad 1$.

Take logarithm of Eq.(6), the nonlinear model becomes linear:

$$Z=\gamma_0+\gamma_1\ln X_1+\gamma_2\ln X_2+\ldots+\gamma_p\ln X_p+\tau, \tag{8}$$

where *Z* follows a normal distribution with mean and variance as

$$E(Z)=\gamma_0+\gamma_1\ln X_1+\gamma_2\ln X_2+\ldots+\gamma_p\ln X_p+\mu, \text{ and} Var(Z)=\sigma_M^2. \tag{9}$$

It should be noted that $E(\tau)\neq 0$, then OLS cannot be applied directly. To make the random error term has zero mean, let $\gamma_0^{¿}$= $\gamma_0+\mu$, then the exponential regression model in Eq.(8) becomes

$$Z = \gamma_0 + \gamma_1 \ln X_1 + \gamma_2 \ln X_2 + \ldots + \gamma_p \ln X_p + \tau^*, \# \qquad (10)$$

where $\tau \sim N\left(0, \sigma_M^2\right)$. Then OLS becomes applicable. To simplify estimation of coefficients, the independent variables are centralized. For *i= 1,..., n*, the linear regression model in Eq.(10) can be written as

$$Z_i = \gamma_0{}' + \gamma_1\left(\ln X_1 - \overline{\ln X_1}\right) + \gamma_2\left(\ln X_2 - \overline{\ln X_2}\right) + \ldots + \gamma_p\left(\ln X_p - \overline{\ln X_p}\right) + \tau^*, \qquad (11)$$

where $\overline{\ln X_k} = \frac{1}{n}\sum_{i=1}^{n} \ln X_{ki}$, for *k*= 1,. . . , *p*,

$\gamma_0{}' = \gamma_0 \quad \overline{\ln X_1} \ldots - \overline{\ln X_p}$, and *n* is the sample size. In matrix

notation, the OLS estimators for coefficients in Eq.(11]) are

$$\hat{\gamma}_0 = \bar{Z} \; and \; \hat{\gamma}_{-0} = S^{-1}T \qquad (12)$$

Where

$$\underset{p\times 1}{\hat{\gamma}_0} = (\hat{\gamma} \ldots \hat{\gamma}_p)^T; \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i; \underset{p\times 1}{s} = (S_{ij}), S_{ij} = \begin{cases} \sum_{i=1}^{n}(In\frac{X_{ik}}{In} - In\bar{X}i) & if\, i=j \\ \sum_{k=1}^{n} Z_i\,(In\,X_{ik-In X_i})\,(In\,x_{Jk} - in\,X_j) & if\, i \neq j \end{cases}; \underset{p\times 1}{T} = (T_i),$$

$$T_i = \sum_{k=}^{n}\left(\ln X_{ik} - \overline{\ln X_i}\right)(Z_k - Z) \; and \; i, j = 1, \ldots, p.$$

$$The \; OLS \; estimator \; for \; \gamma \; is$$

$$\hat{\gamma}^* = \bar{Z} - \overline{\ln X_1} \ldots - \overline{\ln X_p} \qquad (13)$$

Since $\gamma_0 = \gamma^*{}_0 - \mu = -\sigma_M^2/2,$ to estimate γ_0 it is necessary to fin the estimation for σ_M^2 The unbiased estimator for σ_M^2 is [20]

$$\sigma_M^2 = \sum_{i=1}^{n} \frac{(Z_k - \hat{Z}_k)}{n-(p+1)}, \qquad (14)$$

Where $\hat{Z}_k = \hat{\gamma}^*_0 + \hat{\gamma}_1 X_1 + \hat{\gamma}_2 X_2 + \ldots + + \gamma_p X_p$ Then the estimator for $\gamma 0$ is

$$\hat{\gamma}_0 = \hat{\gamma}^*_0 - \hat{\mu} = \gamma_0^* + \frac{\sigma_M^2}{2} \qquad (15)$$

The fitted value for *Z* using Eq.(8) is

$$\hat{Z} = \hat{\gamma}^*_0 + \hat{\gamma}_1 \ln X_1 + \hat{\gamma}_2 \ln X_2 + \ldots + \hat{\gamma}_p \ln X_p = \hat{\gamma}_0 + \hat{\gamma}_1 \ln X_1 + \gamma_2 \ln X_2 + \ldots + \hat{\gamma}_p \ln X_p + \mu. \qquad (16)$$

The estimators for coefficients in Eq.(6) are $\gamma_0^*, \hat{\gamma}_1, \ldots, \hat{\gamma}_p$ and the composite index in Eq.(4) under multiplicative model becomes

$$\hat{C}_M = e^{\hat{\gamma}^*_0} X_1^{\hat{\gamma}_1} X_2^{\hat{\gamma}_2} \ldots X_p^{\hat{\gamma}_p} = e^{\hat{\gamma}_0 + \hat{\mu}} X_1^{\hat{\gamma}_1} X_2^{\hat{\gamma}_2} \ldots X_p^{\hat{\gamma}_p}. \qquad (17)$$

Additive Model

The additive model assumes that the nonlinear regression model has form $e^{\gamma_0} X_1^{\gamma_1} X_2^{\gamma_2} \ldots X_p^{\gamma_p} + \eta$, where ***n*** is the random error term with normal distribution. Denote $\eta \sim N\left(0, \sigma_A^2\right)$. Then the exponential regression model is defined as

$$Z = e^{\gamma_0} X_1^{\gamma_1} X_2^{\gamma_2} \ldots X_p^{\gamma_p} + \eta. \qquad (18)$$

Z will not be transformed as in multiplicative model, since $e^{\gamma_0} X_1^{\gamma_1} X_2^{\gamma_2} \ldots X_p^{\gamma_p} + \eta$ already follows a normal distribution. In nonlinear regression, Gauss-Newton method is a commonly used method to find "best" estimators for coefficients using Taylor expansion. Hartley HO [21] proposed a modified Gauss-Newton method to ensure that every iteration will lead a decrement in sum squares of errors. Graver CA and Boren HE [20] adjusted the modified Gauss-Newton method to make it more efficient in computation. Additionally, the modified Gauss-Newton methods ensures the convergence of estimated coefficients [21].

The computation efficient modified Gauss-Newton method can be summarized as follows [20]. The primary objective is to minimize the sum squares of errors:

$$arg \min_{\gamma_0, \gamma_1, \ldots, \gamma_p} Q(\gamma_0, \gamma_1, \ldots, \gamma_p) = arg \min_{\gamma_0, \gamma_1, \ldots, \gamma_p} \sum_{i=1}^{n}\left(Z_i - e^{\gamma_0} X_{1i}^{\gamma_1} X_{2i}^{\gamma_2} \ldots X_{pi}^{\gamma_p}\right)^2. \qquad (19)$$

Let $\gamma_0^{s\;\;1}, \gamma_1^{s\;\;1}, \ldots, \gamma_p^{s\;\;1}$ denote estimated coefficients in Eq.(18) at the *(S-1)* th iteration. Solve the following *P+1*equations to find the solution for D_0, *D1,..., D_p*: (detailed derivations are in Appendix)

$$\begin{cases} \sum_{i=1}^{n}\left(e^{\gamma_0^{s1}} X_{1i}^{\gamma_1^{s1}} \ldots X_{1i}^{\gamma_1^{s1}}\right)^2 \left(D_0 + \sum_{j=1}^{p} \ln X_{ji} D_j\right) = \sum_{i=1}^{n}\left(Z_i e^{\gamma_0^{s1}} X_{1i}^{\gamma_1^{s1}} \ldots X_{1i}^{\gamma_1^{s1}}\right) e^{\gamma_0^{s1}} X_{1i}^{\gamma_1^{s1}} \ldots X_{1i}^{\gamma_1^{s1}} \\ \sum_{i=1}^{n}\left(e^{\gamma_0^{s1}} X_{1i}^{\gamma_1^{s1}} \ldots X_{1i}^{\gamma_1^{s1}}\right)^2 \left(D_0 \ln X_{ki} + \sum_{j=1}^{p} \ln X_{ki} \ln X_{ji} D_j\right) = \sum_{i=1}^{n}\left(Z_i - e^{\gamma_0^{s1}} X_{1i}^{\gamma_1^{s1}} \ldots X_{1i}^{\gamma_1^{s1}}\right) e^{\gamma_0^{s1}} X_{1i}^{\gamma_1^{s1}} \ldots X_{1i}^{\gamma_1^{s1}} \ln X_{ki} \end{cases} \qquad (20)$$

where *K= 1,..,p.* Then estimated coefficients $\gamma_0^s, \gamma_1^s, \ldots, \gamma_p^s$ at ***S*** th iteration can be updated using the following algorithm:

$k = 1, \ldots, p.\; Then\; estimated\; coefficients\; \gamma_0^s, \gamma_1^s, \ldots, \gamma_p^s\; at\; sth$ iteration can be updated using the following algorithm:

Algorithm 1: Modified Gauss-Newton Algorithm

```
Data: $X_1, \ldots, X_p > 0$ and $Z$
  Result: Estimators of coefficients $(\gamma_0, \gamma_1 \ldots \gamma_p)$ in Eq. (18)
  $(\gamma_0^0, \gamma_1^0, \ldots, \gamma_p^0), S = 1 \; and \; N = 1000;$
  While $s < N$ do
   $solve \; Eq.(20) \; for D_0, D_1, \ldots, D_p;$
  for $l \in \{0, 0.1, 0.2, \ldots, 1.0\}$ do $Q_l \leftarrow Q(\gamma_0^s + l D_0, \gamma_1^s + l D_1 \ldots, \gamma_p^s + l D_p);$
  $l_s \leftarrow arg \; min \, Q_l;$
  $c = 0.1;$
  while $l_s = 0$ & max $l < 10^{-20}$ do
  for $l \in \{c \cdot 0, c \cdot 0.1, c \cdot 0.2 \ldots, c \cdot 1.0\}$ do $Q_l \leftarrow Q(\gamma_0^s + l D_0, \gamma_1^s + l D_1 \ldots, \gamma_p^s + l D_p);$
  $l_s \leftarrow arg \; min \, Q_l$
  $c = c \cdot 0.1;$
  if $l^s \neq 0$ then return $l_s$
  end
  $\gamma_k^{s+1} \leftarrow \gamma_k^s + l^s D_k \; for \; k = 0, 1, \ldots, P;$
  if $\left|\gamma_k^{s+1} - \gamma_k^s\right| < 10^{-5} \; for \; k = 0, 1, \ldots, P$ then return $(\gamma_0^s, \gamma_1^s, \ldots, \gamma_p^s)$;
  $S = S + 1;$
  end
```

Let $\gamma_0^s, \gamma_1^s, \ldots, \gamma_p^s$ denoted the estimated coefficients for nonlinear regression model in Eq.(18) estimated by the modified Gauss-Newton method. The composite index in Eq.(4) under additive model becomes

$$C_A = e^{\gamma_0} X_1^{\gamma_1} X_2^{\gamma_2} \ldots X_p^{\gamma_p}. \qquad (21)$$

Evaluation of Composite Index

One major application of composite index is to alleviate multi-collinearity in regression model. An ideal composite index should yield estimated coefficients with smaller standard error while maintaining the prediction accuracy. Then the composite index can be evaluated in terms of the predicted values of the regression model and the statistical power while testing for coefficients, in addition to standard errors.

Let $\hat{\gamma}°$ denote the prediction from original regression model in Eq.(1). Let $\hat{\gamma}_m$ and $\hat{\gamma}_a$ denoted the predictions from regression model with composite index using multiplicative and additive model, respectively. The following statistics can be used to compare the fitted values using regression models with and without composite index:

Absolute difference: $D_M = \left|\hat{Y}_M - \hat{Y}_0\right) and \; D_A = \left|\hat{Y}_A - \hat{Y}_0\right)$

Relative difference: $R_M = \left|\frac{\hat{Y}_M - \hat{Y}_0}{\hat{Y}_0}\right) and \; R_A = \left|\frac{\hat{Y}_A - \hat{Y}_0}{\hat{Y}_0}\right)$

In addition, the composite index can be evaluated by studying the statistical power of testing for the significance of either the composite index or corresponding highly correlated predictors. For original regression model, the following hypotheses may be tested:

$$H_0 : \beta_1 = ... = \beta_p = 0 \text{ vs } H_1 : \exists\beta_i \neq 0 \text{ for } i = 1,..., p. \tag{22}$$

The partial F test may be applied to test for the hypotheses. Let *A={X*$_{p+1}$*,…,X*$_m$*}* denote the independent variables in the reduced model and the number of variables in A *is w = mp.* Let *B={X*$_{p+1}$*,…,X*$_m$*}* denote the independent variables that we want to test for, and the number of variables in *B* is *u=p*. Under H$_0$, the F statistic is *A={x*$_{p+1}$*,…,X*$_m$*}*

$$F = \frac{\text{SSR(full)} - \text{SSR(reduced)}/u}{\text{SSE(full)}/v} = \frac{\text{SSR(full)} - \text{SSR(reduced)}}{\text{SST} - \text{SSR(full)}} \times \frac{v}{u}, \tag{23}$$

Where *V= n u w* 1, and F follows F distribution with degree of freedom u and v. Reject H$_0$, if $F > c_{\acute{a}}$; otherwise, fail to reject H$_0$, where $c_{\acute{a}} = F_{1\ \acute{a}}^{\ 1}(u\ v\ 0{,}0)$ is the $(1\quad \acute{a})\%$ quantile of F distribution, and $\acute{a}$ is the significance level. The first term of Eq.(23) can be treated as the measure of effect size (the signal-to-noise ratio) and the second term carries information about the sample size and number of parameters [22].Specifically, the effect size is defined as

$$f^2 = \frac{\text{SSR(full)} - \text{SSR(reduced)}}{\text{SST} - \text{SSR(full)}} = \frac{\frac{\text{SSR(full)}}{\text{SST}} - \frac{\text{SSR(reduce)}}{\text{SST}}}{1 - \frac{\text{SSR(full)}}{\text{SST}}}, \tag{24}$$

where$SSR=\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{Y}\right)^2 SSE=\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{Y}\right) SSE=\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{Y}\right)$. It should be noted that SST only depends on the observed data not the model. The correlations between variables in *A,B* and $\tilde{a}$ can be written as

$$R^2_{Y;A,B} = \frac{\text{SSR(full)}}{\text{SST}}, \text{and} R^2_{Y;A} = \frac{\text{SSR(reduce)}}{\text{SST}} \tag{25}$$

Thus, the effect size can be expressed in terms of correlations:

$$f^2 = \frac{R^2_{Y;A,B} - R^2_{Y;A}}{1 - R^2_{Y;A,B}}. \tag{26}$$

Under *H*$_1$, the *F* statistic follows non-central F distribution. Given no centrality parameter *L*, the statistical power is

$$\text{power} = 1 \quad \Psi_{u,v;L}(F), \tag{27}$$

where $\varnothing_{u\ v\ L}(.)$ is the cumulative density function of F distribution with degree of freedom *u* and *v*, and no centrality parameter L ; and F is the observed test statistic in Eq.(23). Using Laubscher's square root normal approximation of noncentral F distribution, the statistical power can be approximated as [22]

$$\text{power} \approx \Phi\left(\frac{\sqrt{2(u+L) \quad \frac{u+2L}{u+L}} \quad \sqrt{(2v \quad 1)\frac{uc_\alpha}{v}}}{\sqrt{\frac{uc_\alpha}{v} + \frac{u+2L}{u+L}}}\right). \tag{28}$$

The non-centrality parameter L and effect size F^2 has the following relation:

$$L = f^v = \frac{R^2_{Y;A,B} - R^2_{Y;A}}{1 - R^2_{Y;A,B}} \times (n - u - w - 1). \tag{29}$$

For regression model using the proposed composite index, the hypotheses in Eq.(22) become

$$H_0 : \beta_c = 0 \text{ vs } H_1 : \beta_c \neq 0. \tag{30}$$

Let B={c} denote the composite index and the number of variables in *B is u=1* . The power of F test (or t test) can be determined similarly by Eq.(27) and Eq.(28).

Simulation

To investigate the performance of the composite index using multiplicative and additive models, a simulation study is conducted to evaluate the probability of the composite index and corresponding highly correlated predictors having consistent statistical inference result, the predicted values from regression model, estimated coefficient standard error, and statistical power of partial F test. Assuming there are four variables in the linear regression model; {X$_1$, X$_2$} is the high-correlation cluster, with correlation denoted as ρ_{high} , {X$_3$, X$_4$} is the low-correlation cluster, with correlation denoted as ρ_{low}, and the between cluster correlation is denoted as ρ_{between}. Specifically, we have

$\rho_{12} = \rho_{\text{high}}$, $\rho_{34} = \rho_{\text{low}}$ and $\rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = \rho_{\text{between}}$.

The simulation contains 1000 runs, and the process is summarized as follows

Generate random samples for *X*$_1$*'*, *X*$_2$*'*, *X*$_3$*'*, *X*$_4$*'* from multivariate normal distribution with mean vector $\grave{\imath}$, standard deviation (standard deviation for all *X*' s are assumed to be the same), and correlation matrix Ω (Ω is determined by ρ_{high} , ρ_{low} , ρ_{between} and In this simulation

$\mu = (\mu_1 \quad \mu_2 \quad \mu_3 \quad \mu_4)^T = (0 \quad 0 \quad 0 \quad 0)^T$ and $\sigma = 1$.

Take exponent for all generated samples for *X*'s to make them satisfy the assumption of being positive for pre-defined exponential model. Let $X_i = e^{X_i}$ (i=1,...,4) denote the variables after taking exponent. It should be noted that the proposed approaches only require the high-correlation cluster variables to be positive. All independent variables are taken exponent in the simulation for simplicity.

The dependent variable is generated by

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon,\ \textit{Where}\ \ \varepsilon \sim N(0,1)$

Fit regression model in Eq.(1) and (5) and evaluate composite index.

Let $\rho_{\text{between}} = 0.05$ to make the data have low between-cluster correlation. ρ_{high} takes values 0.7 (moderate high),0.80 (high) and 0.90 (very high). ρ_{low} takes values 0.1 (very low), 0.2 (low) and 0.3 (moderate low). Let $\beta_0 = \beta_3 = \beta_4 = 1$, and sample size *n=50.* The significance level is $\alpha = 0.05$.

(Table 1) presents the conditional probabilities for composite index having significant coefficients given either X$_1$ *or X*$_2$ with significant coefficient. Here, three situations are considered, including $\beta_1 = \beta_2 = 3$ when *X*$_1$ *and X*$_2$ have great impact on the outcome, $\beta_1 = \beta_2 = 0.3$ when *X*$_1$ *and X*$_2$ have small impact on the outcome, and $\beta_1 = 3, \beta_2 = 0.3$ when *X*$_1$ has great impact and *X*$_2$ has small impact. For both multiplicative and additive composite index, if the original highly correlated predictors are significant, the composite index will almost always be significant, regardless of whether the impact on the outcome is big or small.

On the other hand, the conditional probabilities for or being significant given the composite index is significant are shown in (Table 2). When , the conditional probabilities for or being significant are always . It suggests that when the impact of original predictors is highly significant, original predictors will still yield significant conclusions even though multi-collinearity exists in the model. However, when the true coefficients decrease to , the conditional probabilities for or being significant decreases dramatically, especially when and are highly correlated. Specifically, when the within-cluster correlation is the conditional probabilities for (or) become less than. In this way, multi-collinearity can greatly influence the conclusion of the regression model. In summary, the proposed multiplicative and additive composite index will be significant if either or is significant, i.e., they have better performances in informing disease status and detecting treatment effect than and in a multi-collinear regression model.

As mentioned in Section 4, a valid composite index is expected to yield non-inferior predictions than regression model without composite index. (Table 3) presents the absolute and relative difference of fitted values for regression model with and without composite index. Due to having fewer predictors in the model, both multiplicative and additive composite index tend to provide predictions different from the predictive model without considering composite index. And this difference decreases when and/or have less impact on the outcome variable. Comparing absolute differences, additive composite index has better performance than multiplicative composite index. Fixed , absolute differences increase when increases. Similar findings are detected while fixing. As for relative differences in predicted values, multiplicative composite index also leads to higher differences than additive composite index. For additive composite index, the relative difference of additive composite index is much closer to compared with multiplicative composite index.

Table 1. Conditional probabilities for composite index to be significant given original predictor is significant.

		Multiplicative composite index			Additive composite index		
ρ *High*	ρ *low*	$P_{X_1}\|C^1$	$P_C\|X^2$ ^{2}	$P_C\|X_1 or\ X_2$ ^{3}	$P_C\|X_1$	$P_C\|X_2$	$P_C\|X_1 or X_2$
Case 1 : $\beta_1 = \beta_2 = 3$							
0.7	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	0.999	0.999	0.999	1.000	1.000	1.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000
0.8	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	0.999	0.999	0.999	1.000	1.000	1.000
	0.3	0.998	0.998	0.998	1.000	1.000	1.000
0.9	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	0.998	0.998	0.998	1.000	1.000	1.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000
Case 2: $\beta_1 = \beta_2 = 0.\ 3$							
0.7	0.1	0.999	1	0.999	1.000	1.000	1.000
	0.2	0.999	1	0.999	1.000	1.000	1.000
	0.3	1.000	1	1.000	1.000	1.000	1.000
0.8	0.1	1.000	1	1.000	1.000	1.000	1.000
	0.2	1.000	1	1.000	1.000	1.000	1.000
	0.3	1.000	1	1.000	1.000	1.000	1.000
0.9	0.1	1.000	1	1.000	1.000	1.000	1.000
	0.2	1.000	1	1.000	1.000	1.000	1.000
	0.3	1.000	1	1.000	1.000	1.000	1.000
Case 3: $\beta_1 = 3, \beta_{2 =}\ 0.\ 3$							
0.7	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	1.000	1.000	1.000	1.000
	0.3	0.999	1.000	0.999	1.000	1.000	1.000
0.8	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	1.000	1.000	1.000	1.000
	0.3	0.999	0.998	0.999	1.000	1.000	1.000
0.9	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	1.000	1.000	1.000	1.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000

$^1P_C|X_1 = P(C\ significant|X_1\ significant);$

$^2P_C|X_2 = P(C\ significant|X_1\ significant);$

$^3P_C|X_1 or\ \ X_2 = P(C\ significant|X_1\ or\ X_2 significant),$

Where stands for either multiplicative composite index or additive composite index.

Table 2. Conditional probabilities for original predictors to be significant given composite index is significant.

		Multiplicative composite index			Additive composite index		
ρ *High*	ρ *low*	$P_{X_1}\|C^1$	$P_{X2}\|C^2$	$P_{X_1}\ or\ _{X2}\|C^3$	$P_{X_1}\|C$	$P_{X_2}\|C$	$P_{X_1}\ or\ _{X2}\|C$
Case 1 : $\beta_1 = \beta_2 = 3$							
0.7	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	1.000	1.000	1.000	1.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000
0.8	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	1.000	1.000	1.000	1.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000
0.9	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	1.000	1.000	1.000	1.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000
Case 2: $\beta_1 = \beta_2$=0. 3							
0.7	0.1	0.768	0.718	0.960	0.765	0.715	0.956
	0.2	0.729	0.734	0.959	0.727	0.731	0.957
	0.3	0.731	0.754	0.954	0.731	0.754	0.954
0.8	0.1	0.634	0.658	0.919	0.634	0.658	0.919
	0.2	0.632	0.676	0.911	0.632	0.676	0.911
	0.3	0.661	0.660	0.924	0.661	0.660	0.923

0.9	0.1	0.443	0.439	0.736	0.443	0.439	0.736
	0.2	0.419	0.463	0.739	0.418	0.462	0.737
	0.3	0.432	0.438	0.747	0.432	0.438	0.748
			Case 3: $\beta_1 = 3$, $\beta_2 = 0.3$				
0.7	0.1	1.000	0.771	1.000	1.000	0.771	1.000
	0.2	1.000	0.732	1.000	1.000	0.732	1.000
	0.3	1.000	0.756	1.000	1.000	0.755	1.000
0.8	0.1	1.000	0.658	1.000	1.000	0.658	1.000
	0.2	1.000	0.631	1.000	1.000	0.631	1.000
	0.3	1.000	0.627	1.000	1.000	0.627	1.000
0.9	0.1	1.000	0.430	1.000	1.000	0.430	1.000
	0.2	1.000	0.447	1.000	1.000	0.447	1.000
	0.3	1.000	0.461	1.000	1.000	0.461	1.000

[1] $Note: P_{X_1}|C = P(X_1\ significant|C\ significant);$

[2] $P_{X_2}|C = P(X_1\ significant|C\ significant);$

[2] $P_{X_1} or_{X_2|C} = P(X_1\ significant|C\ significant);$

Table 3. Evaluation of composite index in terms of fitted values of regression model.

ρ_{High}	ρ_{low}	$\bar{D}_M$[1]	$\bar{D}_A$[1]	$\bar{R}_M$[2]	$\bar{R}_A$[2]
		Case 1 : $\beta_1 = \beta_2 = 3$			
0.7	0.1	5.299	0.604	0.523	0.044
	0.2	5.324	0.606	0.525	0.044
	0.3	5.377	0.629	0.53X3	0.045
0.8	0.1	5.416	0.405	0.538	0.030
	0.2	5.448	0.416	0.540	0.030
	0.3	5.519	0.425	0.553	0.031
0.9	0.1	5.532	0.207	0.553	0.015
	0.2	5.612	0.207	0.561	0.015
	0.3	5.604	0.208	0.566	0.015
		Case 2: $\beta_1 = \beta_2 = 0.3$			
0.7	0.1	0.125	0.057	0.028	0.012
	0.2	0.129	0.056	0.029	0.012
	0.3	0.132	0.057	0.030	0.012
0.8	0.1	0.127	0.036	0.028	0.008
	0.2	0.126	0.037	0.029	0.008
	0.3	0.123	0.036	0.028	0.008
0.9	0.1	0.124	0.018	0.028	0.004
	0.2	0.124	0.018	0.028	0.004
	0.3	0.118	0.017	0.027	0.004
		Case 3: $\beta_1 = 3$, $\beta_2 = 0.3$			
0.7	0.1	2.861	0.110	0.377	0.012
	0.2	2.847	0.106	0.378	0.012
	0.3	2.830	0.107	0.380	0.012
0.8	0.1	2.835	0.072	0.375	0.008
	0.2	2.855	0.070	0.383	0.008
	0.3	2.845	0.072	0.384	0.008
0.9	0.1	2.858	0.037	0.381	0.004
	0.2	2.847	0.038	0.383	0.004
	0.3	2.842	0.038	0.384	0.004

D_M and D_M are sample average of D_M and D_A , respectively;
R_M and R_A are sample average of R_M and R_A , respectively.

Under the same parameter setting as (Table 3) and (Table 4) presents the evaluation findings in terms of standard errors. The empirical probability of composite index has smaller estimated coefficient standard error is studied. The estimated coefficient standard error of composite index is compared to both the minimum and maximum of the standard error for estimated coefficients of and . For multiplicative composite index, the probability of having a lower standard error is 1 under all settings, suggesting that the multiplicative composite index can always successfully alleviate the problem of having large standard error due to multi-collinearity. As for additive composite index, the probability for it to have smaller standard error than at least one of and is not exactly , however, the probability relatively large, especially when the within-cluster correlation is very high. And the additive composite index is less likely to have a smaller standard error than all predictors compared with multiplicative composite index.

Table 4. Evaluation of composite index in terms of standard error of estimated coefficient.

		Multiplicative composite index		Additive composite index	
ρ_{High}	ρ_{low}	P_{min}	P_{max}	P_{min}	P_{max}
Case 1 : $\beta_1 = \beta_2 = 3$					
0.7	0.1	1.000	1.000	0.602	1.000
	0.2	1.000	1.000	0.294	1.000
	0.3	1.000	1.000	0.591	1.000
0.8	0.1	1.000	1.000	0.791	1.000
	0.2	1.000	1.000	0.813	1.000
	0.3	1.000	1.000	0.790	1.000
0.9	0.1	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	0.998	1.000
	0.3	1.000	1.000	0.995	1.000
Case 2: $\beta_1 = \beta_2 = 0.3$					
0.7	0.1	0.321	0.999	0.001	0.985
	0.2	0.223	0.996	0.000	0.967
	0.3	0.166	0.998	0.000	0.972
0.8	0.1	0.282	1.000	0.000	0.991
	0.2	0.238	1.000	0.000	0.981
	0.3	0.360	0.999	0.003	0.993
0.9	0.1	0.589	1.000	0.010	0.999
	0.2	0.323	0.999	0.000	0.998
	0.3	0.397	1.000	0.001	0.999
Case 3: $\beta_1 = 3$, $\beta_2 = 0.3$					
0.7	0.1	1.000	1.000	0.113	1.000
	0.2	1.000	1.000	0.594	1.000
	0.3	1.000	1.000	0.613	1.000
0.8	0.1	1.000	1.000	0.800	1.000
	0.2	1.000	1.000	0.730	1.000
	0.3	1.000	1.000	0.635	1.000
0.9	0.1	1.000	1.000	0.938	1.000
	0.2	1.000	1.000	0.976	1.000
	0.3	1.000	1.000	0.977	1.000

Note: $P_{min} = P\left(\mathrm{SE}_C < min\{\mathrm{SE}_{X_1}, \mathrm{SE}_{X_2}\}\right)$ and $P_{max} = P\left(\mathrm{SE}_C < max\{\mathrm{SE}_{X_1}, \mathrm{SE}_{X_2}\}\right)$, where $\mathbf{SE_C}$ is the standard error of the estimated coefficient of the composite index using either multiplicative model or additive model, SE_{X1} and SE_{X2} are standard errors of the estimated coefficients of X_1 and X_2 in regression model without using composite index.

Summarizing findings for informing disease status and/or detecting treatment effects, based on results presented in Table 1 and 2, both multiplicative and additive composite index have very good performance. Specifically, if the original highly correlated predictors are significant, the proposed composite index will always be significant. Additionally, when and/or are not significant due to multi-collinearity, the proposed composite index can successfully inform disease status. To compare the performance between multiplicative and additive composite index, the estimated coefficient standard error and fitted values of regression model are studied. It turns out that additive composite index tends to make more accurate predictions (as shown in Table 4), and multiplicative composite index tends to have smaller standard error (as shown in (Table 4)).

Additionally, the empirical statistical power of partial F test with and without using composite index is shown in (Figure 1). Specifically, it presents the empirical statistical power to test for hypotheses in Eq.([eq18]) when the statistical power while using any composite index is always higher than situation without using composite index. The difference in statistical power decreases when sample size increases. If the sample size is larger than , under the specific parameter setting, differences in empirical powers become negligible. Additionally, the statistical power of using additive composite index is higher than multiplicative composite index when sample size is small in most cases. Thus, both multiplicative and additive composite index lead to higher statistical power of partial F test than situations ignoring multi-collinearity in the regression model.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Derivation for modified Gauss-Newton method

Under the assumption for $\eta \sim N\left(0, \sigma_A^2\right)$, the additive model is defined as

$$Z = e^{\gamma_0} X_1^{\gamma_1} X_2^{\gamma_2} \ldots X_p^{\gamma_p} + \eta := f(X;\gamma) + \eta,$$

where $\gamma = \left(\gamma_0 \quad \gamma_1 \quad \ldots \quad \gamma_p\right)^T$. For every observation, $Z_i = f\left(X_i;\gamma\right)$ for *i=1,...,n*. Similar to OLS, the "best" estimates for coefficients is defined as the estimates that minimize the sum squares of errors as Eq.(19):

$$arg \min_{\gamma} Q(\gamma) = arg \min_{\gamma} \sum_{i=1}^{n} \left(Z_i \quad f\left(X_i;\gamma\right)\right)^2. \qquad \text{(A1)}$$

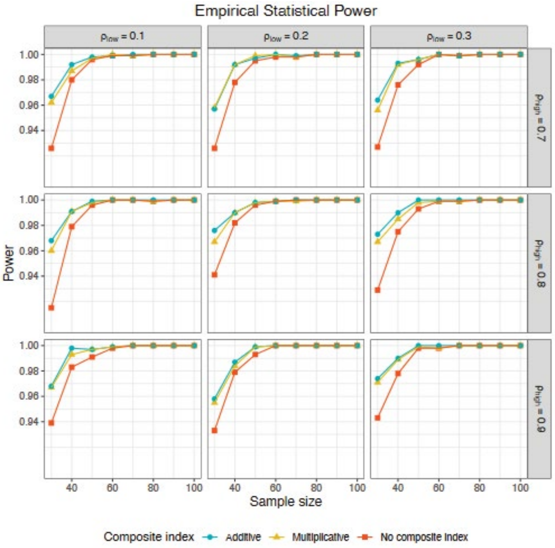

Figure 1. Statistical power of partial F test with and without using composite index. Under the specific parameter setting in thesimulation, the sample size cannot be smaller than 30, otherwise computational singularity problem will occur in estimating coefficients for additive model.

To find of $argminQ(\gamma)$, one may consider finding the solution to $\frac{\partial Q(\gamma)}{\partial \gamma_k} = 0$ for $k = 0,1,\ldots,p$. However, since $f(X_i;\gamma)$ is a nonlinear function, it is quite difficult to take first derivative of $Q(\gamma)$ directly.

Utilizing Taylor expansion $f(X_i;\gamma)$ can be approximated using a linear function. Specifically, the first order Taylor expansion of $f(X_i;\gamma)$ at $\gamma = \gamma^0$ is

$$f(X_i,\gamma)f(X_i,\gamma^0)+\sum_{j=0}^{p}\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_j}\right)_{\gamma=\gamma^0}(\gamma_j-\gamma_j^0). \quad \text{(A2)}$$

Then $Q(\gamma)$ in Eq.(A1) can be written as

$$Q(\gamma)=\sum_{i=1}^{n}\left(Z_i-f(X_i,\gamma^0)-\sum_{j=0}^{p}\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_j}\right)_{\gamma=\gamma^0}(\gamma_j-\gamma_j^0)\right)^2 \quad \text{(A3)}$$

Then the first derivative of $Q(\gamma)$ with respect to γ_k, for *K = 0,1,…,p* can be written as

$$Q_k(\gamma)=\sum_{i=1}^{n}2\left(Z_i-f(X_i,\gamma^0)-\sum_{j=0}^{p}\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_j}\right)_{\gamma=\gamma^0}(\gamma_j-\gamma_j^0)\right)\cdot\left(-\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_j}\right)_{\gamma=\gamma^0}\right). \quad \text{(A4)}$$

Let $f_k(X_i;\gamma_0)$ denote $\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_j}\right)_{\gamma=\gamma^0}$. $Q_k(\gamma)$ Can be further written as

$$Q_k(\gamma)=\sum_{i=1}^{n}2\left(Z_i-f(X_i,\gamma^0)-\sum_{i=1}^{n}f_j(X_i;\gamma^0)(y_{j-}y_j^0)\right).(-f_k(X_i;\gamma^0))$$

$$=-2\sum_{i=1}^{n}(Z_i-f(X_i;\gamma^0)).f_k(X_i;\gamma^0)+2\sum_{i=1}^{n}\sum_{j=0}^{p}f_j(X_i;\gamma^0)f_k(X_i;\gamma^0)(y_{j-}y_j^0). \quad \text{(A5)}$$

Since

$$Q_k(\gamma^0)=\left[\frac{\partial Q(\gamma)}{\partial \gamma_k}\right)_{\gamma=\gamma^0}=-2\sum_{i=1}^{n}(Z_i-f(X_i;\gamma^0))f_k(X_i;\gamma^0), \quad \text{(A6)}$$

By setting the first derivative equals to zero, we have

$$\Leftrightarrow 2\sum_{i=1}^{n}\sum_{j=0}^{p}f_j(X_i;\gamma^0)f_k(X_i;\gamma^0)(y_{j-}y_j^0)=-Q_k(\gamma^0). \quad (A7)$$

Let $D_j = \gamma_j\ \gamma_j^0$, Eq. (A7) can be written as

$$2\sum_{i=1}^{n}\sum_{j=0}^{p}f_j(X_i;\gamma^0)f_k(X_i;\gamma^0)D_j=Q_k(\gamma^0). \quad \text{(A8)}$$

For *K=0 ,1,…,P,*, the first derivatives of $f(X_i,\tilde{a})$ with respect to $\tilde{a}_k$ can be derived as

$$\begin{Bmatrix} f_0(X_i,\gamma^0)=\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_0}\right)_{\gamma=\gamma^0}=\left[e^{\gamma_0}X_{1i}^{\gamma_1}X_{2i}^{\gamma_2}\ldots X_{pi}^{\gamma_p}\right)_{\gamma=\gamma^0}=e^{\gamma_0^0}X_{1i}^{\gamma_1^0}X_{2i}^{\gamma_2^0}\ldots X_{pi}^{\gamma_p^0} & \text{for}k=0 \\ f_k(X_i,\gamma^0)=\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_k}\right)_{\gamma=\gamma^0}=\left[e^{\gamma_0}X_{1i}^{\gamma_1}X_{2i}^{\gamma_2}\ldots X_{pi}^{\gamma_p}\ln X_{ki}\right)_{\gamma=\gamma^0}=e^{\gamma_0^0}X_{1i}^{\gamma_1^0}X_{2i}^{\gamma_2^0}\ldots X_{pi}^{\gamma_p^0}\ln X_{ki} & \text{for}k=1,\ldots,p \end{Bmatrix} (A9)$$

Then the following linear *P +1* equations can be derived:

$$\begin{Bmatrix} \sum_{i=1}^{n}\left(e^{\mu_0^0}X_{1i}^{\mu_1^0}\ldots X_{pi}^{\mu_p^0}\right)^2\left(D_0+\sum_{j=1}^{p}\ln X_{ji}D_j\right)=\sum_{i=1}^{n}\left(Y_i-e^{\mu_0^0}X_{1i}^{\mu_1^0}\ldots X_{pi}^{\mu_p^0}\right)e^{\mu_0^0}X_{1i}^{\mu_1^0}\ldots X_{pi}^{\mu_p^0} \\ \sum_{i=1}^{n}\left(e^{\mu_0^0}X_{1i}^{\mu_1^0}\ldots X_{pi}^{\mu_p^0}\right)^2\left(D_0\ln X_{ki}+\sum_{j=1}^{p}\ln X_{ki}\ln X_{ji}D_j\right)=\sum_{i=1}^{n}\left(Y_i-e^{\mu_0^0}X_{1i}^{\mu_1^0}\ldots X_{pi}^{\mu_p^0}\right)e^{\mu_0^0}X_{1i}^{\mu_1^0}\ldots X_{pi}^{\mu_p^0}\ln X_{ki} \end{Bmatrix},(A10)$$

$Where\ K=1,\ldots,p$

Concluding Remarks

In this research, we aim to propose a valid composite index which can not only alleviate multi-collinearity problem while maintaining predictive capability. In other words, in a regression model, the composite index should have a smaller standard error for the estimated coefficients compared with the original highly correlated predictors, while making accurate predictions. One typical structure for composite index in clinical research considered in this research is the exponential structure. One example of an exponential-type composite index which can inform disease status as well as treatment effect is the BMI. BMI combines weight and height into a single index which can best inform the obesity (disease status) and treatment effect (reduction of obesity) of the participants under study. The advantage of using exponential-type composite index is that it is consistent many well-accepted clinical composite index and easy to interpret, especially when there are fewer highly correlated predictors.

Two methods are proposed in this paper to establish an exponential-type composite index, namely the multiplicative model and the additive model. The multiplicative model takes the error term in a “multiplicative” way, whereas the additive model has an added error term. The composite indices derived from each model are named multiplicative and additive composite index. The “weights” of variables constituting the composite index are determined by the regression coefficients. In the multiplicative model, the error term assumes to follow log-normal distribution. After taking logarithm, the problem of finding the weights becomes estimating coefficients of linear regression. Then OLS method can be applied after some algebraic computation. As for the additive model, the error term assumes to follow a normal distribution. The modified Gauss-Newton method is applied to estimate coefficients. In this way, the composite index can be derived.

According to simulation findings, both multiplicative and additive composite indices can improve the estimated coefficients by having a smaller standard error. Specifically, the multiplicative composite index has a smaller standard error for estimated coefficient with empirical probability of, and the additive composite index has corresponding probability close to comparing the predictive ability, the additive composite index has better performance in terms of both absolute difference and relative difference to the original regression model. Studies on empirical statistical power for the partial F test illustrate that both composite indices have higher statistical power than the regression model with multi-collinearity issue. The difference among statistical powers decreases as sample size increases. Under the specific parameter setting, the difference becomes negligible when the sample size is larger than.

One major assumption of the proposed models is that the predictors used to build a composite index must be strictly positive all the time. In future research, other types of predictors can be considered, for example, binary predictor or predictors with negative values. In either scenario, the exponential-type structure and weight calculation method need to be generalized. In addition to alleviating multi-collinearity in regression models, the idea of composite index can be generalized to form a composite outcome variable to overcome multiple testing problem. Specifically, Chow SC and Patty JL [18] proposed the idea of using therapeutic index to utilize information from all primary study outcomes. Overall, this paper may benefit researchers in alleviating multi-collinearity in linear regression models and establishing an innovative composite index to interpret clinical findings. Based on findings from the paper, the multiplicative composite index is recommended if making inferences on treatment effects is the main objective, and the additive composite index is preferred if researchers are more interested in maintaining predictive ability.

Acknowledgements

There is no conflict of interests.

References

1. Song, Mi-Kyung, Feng-Chang Lin, Sandra E. Ward and Jason P. Fine. "Composite variables: When and how." *Nurs Res* 62 (2013): 45.
2. Andrade, Chittaranjan. "Z scores, standard scores, and composite test scores explained." *Indian J Psychol Med* 43 (2021): 555-557.
3. Petrini, Juliana, Raphael Antonio Prado Dias, Simone Fernanda Nedel Pertile and Joanir Pereira Eler, et al. "Degree of multicollinearity and variables involved in linear dependence in additive-dominant models." *Pesqui Agropecu Bras* 47 (2012): 1743-1750.
4. Allen, Michael Patrick. "The problem of multicollinearity." *Understanding regression analysis* (1997): 176-180.
5. Aletaha, D. and J. Smolen. "The simplified disease activity index (SDAI) and the clinical disease activity index (CDAI): A review of their sefulness and validity in rheumatoid arthritis." *Clin Exp Rheumatol* 23 (2005): S100.
6. Chang, Peter, Konrad M. Szymanski, Rodney L. Dunn and Jonathan J. Chipman, et al. "Expanded prostate cancer index composite for clinical practice: Development and validation of a practical health related quality of life instrument for use in the routine clinical care of patients with prostate cancer." *J Urol Balt* 186 (2011): 865-872.
7. Dong, Lu, Armando J. Martinez, Daniel J. Buysse and Allison G. Harvey. "A composite measure of sleep health predicts concurrent mental and physical health outcomes in adolescents prone to eveningness." Sleep Health 5 (2019): 166-174.
8. Chao, Yi-Sheng and Chao-Jung Wu. "Principal component-based weighted indices and a framework to evaluate indices: Results from the Medical Expenditure Panel Survey 1996 to 2011." *PloS One* 12 (2017): e0183997.
9. Chao, Yi-Sheng, Chao-Jung Wu, Hsing-Chien Wu and Hui-Ting Hsu, et al. "Critical hierarchical appraisal and reporting tool for composite measures (CHAOS)." *Cureus* 15 (2023).
10. Liu, Chenyu, Xinlian Zhang, Tanya T. Nguyen and Jinyuan Liu, et al. "Partial least squares regression and principal component analysis: Similarity and differences between two popular variable reduction approaches." *Gen Psychiatr* 35 (2022).
11. Lovaglio, Pietro Giorgio and Giorgio Vittadini. "Structural equation models in a redundancy analysis framework with covariates." *Multivariate Behav Res* 49 (2014): 486-501.
12. Gajewski, Byron J. and Nancy Dunton. "Identifying individual changes in performance with composite quality indicators while accounting for regression to the mean." *Med Decis Making* 33 (2013): 396-406.
13. Proust-Lima, Cécile, Viviane Philipps, Jean-François Dartigues and David A. Bennett, et al. "Are latent variable models preferable to composite score approaches when assessing risk factors of change? Evaluation of type-I error and statistical power in longitudinal cognitive studies." *Stat Methods Med Res* 28 (2019): 1942-1957.
14. Barker, David H., Lori AJ Scott-Sheldon, Daniel Gittins Stone and Larry K. Brown. "Using composite scores to summarize adolescent sexual risk behavior: Current state of the science and recommendations." *Arch Sex Behav* 48 (2019): 2305-2320.
15. Paruolo, Paolo, Michaela Saisana and Andrea Saltelli. "Ratings and rankings: Voodoo or science?." *appl stat* 60 (2015): 69-70.
16. Vollmer, Robin T. "The importance of tumor length in needle biopsies of the prostate." *Am J Clin Pathol* 154 (2020): 533-535.
17. Becker, William, Michaela Saisana, Paolo Paruolo and Ine Vandecasteele. "Weights and importance in composite indicators: Closing the gap." *Ecol Indic* 80 (2017): 12-22.
18. Chow, Shein-Chung and Patty J Lee. "Time to revisit endpoint selection in clinical trials." *In Sixth Eurasian Conference on Language and Social Sciences* (ECLSS2019b) 9 (2020): 247–50.
19. Cohen, Jacob. ‘’Statistical power analysis for the behavioral sciences.’’ Academic press (2013).
20. Graver, C. A. and H. E. Boren. ‘’Multivariate logarithmic and exponential regression models’’. Rand Corporation (1967).
21. Hartley, Herman O. "The modified gauss-newton method for the fitting of non-linear regression functions by least squares." Technometrics 3 (1961): 269-280.
22. Vink, Arja Suzanne, Benjamin Neumann, Krystien VV Lieve and Moritz F. Sinner, et al. "Determination and interpretation of the QT interval: Comprehensive analysis of a large cohort of long QT syndrome patients and controls." Circ 138 (2018): 2345-2358.

How to cite this article: Wang, Peijin. “Development of Composite Index in Clinical Research.” *J Biom Biosta* 14 (2023): 155