PLOS ONE

RESEARCH ARTICLE

On tests of treatment-covariate interactions: An illustration of appropriate power and sample size calculations

Gwowen Shieh*

Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan

* gwshieh@mail.nctu.edu.tw

Citation: Shieh G (2017) On tests of treatment-covariate interactions: An illustration of appropriate power and sample size calculations. PLoS ONE 12 (5): e0177682. https://doi.org/10.1371/journal.pone.0177682

Editor: Jake Olivier, University of New South Wales, AUSTRALIA

Received: January 22, 2017

Accepted: April 30, 2017

Published: May 17, 2017

Data Availability Statement: Data are from the book of Fleiss, J. L. (2011). Design and analysis of clinical experiments (Vol. 73). New York, NY: Wiley.

Funding: This study was funded by Ministry of Science and Technology of Taiwan with the grant: MOST 105-2410-H-009 -035 -MY2.

Competing interests: The author has declared that no competing interests exist.

Abstract

The appraisals of treatment-covariate interaction have theoretical and substantial implications in all scientific fields. Methodologically, the detection of interaction between categorical treatment levels and continuous covariate variables is analogous to the homogeneity of regression slopes test in the context of ANCOVA. A fundamental assumption of ANCOVA is that the regression slopes associating the response variable with the covariate variable are presumed constant across treatment groups. The validity of homogeneous regression slopes accordingly is the most essential concern in traditional ANCOVA and inevitably determines the practical usefulness of research findings. In view of the limited results in current literature, this article aims to present power and sample size procedures for tests of heterogeneity between two regression slopes with particular emphasis on the stochastic feature of covariate variables. Theoretical implications and numerical investigations are presented to explicate the utility and advantage for accommodating covariate properties. The exact approach has the distinct feature of accommodating the full distributional properties of normal covariates whereas the simplified approximate methods only utilize the partial information of covariate variances. According to the overall accuracy and robustness, the exact approach is recommended over the approximate methods as a reliable tool in practical applications. The suggested power and sample size calculations can be implemented with the supplemental SAS and R programs.

Introduction

The existence of interactive phenomena between predictor variables on the response variable is an essential issue in all scientific studies. The detection of interactions between categorical treatment levels and continuous covariate variables is equivalent to the test of homogeneity of regression slopes test in ANCOVA designs. Notably, ANCOVA represents a constructive synthesis of analysis of variance and multiple linear regression to account for the relationship between the response variable and the concomitant or covariate variables in treatment comparisons. In addition to the fundamental assumptions of independence, normality, and constant variance, the within-group regression coefficients of the criterion variable on the

covariate variable are presumed to be equal in ANCOVA. Violation of the ANCOVA assumptions has been the target of attention in the literature such as Glass, Peckham, and Sanders [1] and Harwell [2]. Naturally, the actual significance level and power of the regular test for treatment effects can be distorted to some extent under nonparallel regression settings. Hence, the validity of heterogeneity regression slopes plays a crucial role in applying the traditional ANCOVA or generalized alternatives. As a general guideline, a test for nonparallel regression lines is required as the preliminary procedure for use of traditional ANCOVA. If the test for heterogeneity of regression slopes is significant, then it suggests that the standard ANCOVA is no longer an appropriate technique. Accordingly, Fleiss [3], Huitema [4], and Maxwell and Delaney [5] provide comprehensive exposition and general strategy under heterogeneity of regression.

The statistical perspectives and appropriate strategies of covariate selection are presented in Hauck, Anderson, and Marcus [6], Hernandez, Steyerberg, and Habbema [7], Pocock et al. [8], Raab, Day, and Sales [9], and references therein. Moreover, the impact of omitted covariates on the statistical inferences has been demonstrated in Hauck et al. [10], Gail, Wieand, and Plantadosi [11], and Negassa and Hanley [12]. However, there is no related exploration about the direct consequence of excluding covariate characteristics in power and sample size calculations. In view of the potential applicability in practice, this article focuses on the most fundamental ANCOVA designs for two treatment groups and a single covariate. For the purposes of planning research designs and validating crucial interactions, power and sample size procedures were considered in Dupont and Plummer [13]. Their formula is very attractive from a computational standpoint and has been implemented in statistical packages. However, it is important to note that the particular method involves several convenient approximations including the use of a shifted t distribution for a noncentral t distribution and the substitution of fixed parameters for random covariates. The inherent nature and implications of accuracy were not addressed in Dupont and Plummer [13]. Accordingly, the existing illustrations were not detailed enough to elucidate the potential deficiency of their approximate technique. Because of the limited results in the literature, the current article aims to contribute to the development of power and sample size methodology for the tests for heterogeneity of two regression slopes. The emphasis is placed on the practical situation that not only the values of response variables for each subject are just available after the observations are made, but also the levels of covariate variables cannot be predetermined before data collection.

It is noteworthy that a different and prominent situation of interactive research involves interactions between two continuous covariates. Although the model formulations and test procedures of the interactive analysis are rather similar for the two types of covariate variable combination: continuous by continuous and categorical by continuous, their test statistics and associated distribution properties are considerably different. Therefore, the power and sample size calculations of Shieh [14] for detecting interactions between two continuous variables in multiple regression settings are not appropriate for assessing interactions between grouping and continuous variables within the context of ANCOVA. In a continual effort to support the analytical development and improve the essence of research findings in interaction studies, this investigation updates and expands the previous work of Dupont and Plummer [13] in such a way that the findings not only notify the fundamental deficiency of existing procedure, but also reinforce the usefulness of interaction designs in applications.

The present study has three key aspects. First, to account for the stochastic nature of covariate variables, the covariates are assumed to follow a normal distribution. Both exact and approximate power functions and sample size procedures for detecting heterogeneity of regression slopes are derived. Second, extensive numerical examinations were conducted to examine the deficiency of the approximate methods and the advantage of the exact approach

under a wide range of model settings. The performance and robustness of the described techniques with respect to non-normality of the covariates are also investigated. Third, in view of the limited features of existing software packages, both SAS [15] and R [16] computer algorithms are developed to facilitate the implementation of the suggested power and sample size computations.

Methods

The two-group nonparallel simple linear regression model is of the form

$$Y_{1j} = \beta_{01} + X_{1j}\beta_{11} + \varepsilon_{1j} \text{ and } Y_{2k} = \beta_{02} + X_{2k}\beta_{12} + \varepsilon_{2k}, \quad (1)$$

where ε_{1j} and ε_{2k} are *iid* $N(0,\sigma^2)$ random variables, $j = 1,\ldots, N_1$, and $k = 1,\ldots, N_2$. It is often informative to rewrite the regression model with heterogeneous slopes in Eq 1 as the following interactive multiple regression model using a dummy variable M:

$$Y_i = \beta_{02} + M_i\beta_{0D} + X_i\beta_{12} + M_iX_i\beta_{1D} + \varepsilon_i, i = 1,\ldots,N, N = N_1 + N_2, \quad (2)$$

$$where \ \beta_{0D} = \beta_{01} - \beta_{02}, \beta_{1D} = \beta_{11} - \beta_{12};$$

$$Y_i = Y_{1j}, X_i = X_{1j}, \varepsilon_i = \varepsilon_{1j}, \ and \ M_i = 1 \ if \ i = j, j = 1,\ldots,N_1;$$

$$Y_i = Y_{2k}, X_i = X_{2k}, \varepsilon_i = \varepsilon_{2k}, \ and \ M_i = 0 \ if \ i = N_1 + k, k = 1,\ldots,N_2.$$

Note that a traditional ANCOVA model assumes that the regression slopes are equivalent $\beta_{11} = \beta_{12} = \beta_1$ and it postulates the parallel regression formulation

$$Y_i = \beta_{02} + M_i\beta_{0D} + X_i\beta_1 + \varepsilon_i, i = 1,\ldots,N. \quad (3)$$

Because the strategy and procedure for treatment comparisons differ for the nonparallel and parallel regression frameworks, the equality of covariate regression coefficients is viewed as the most crucial assumption in ANCOVA. Accordingly, a test for heterogeneity of regression slopes is generally required to justify the use of ANCOVA. When the assumption of equal within-group covariate regression coefficients is not tenable, the standard procedures of ANCOVA are no longer appropriate and alternative methods such as Johnson-Neyman and Picked-Point solutions for heterogeneous regression should be adopted. More conceptual and thorough discussions of alternative solutions to traditional ANCOVA can be found in Rogosa [17] and Rutherford [18].

In order to facilitate the detection of heterogeneous regression slopes, this article describes and examines the corresponding procedures for power and sample size determinations. Under the heterogeneous linear model assumption defined in Eq 1, it follows from standard results that the least squares estimators $\hat{\beta}_{11}$ and $\hat{\beta}_{12}$ of slope coefficients β_{11} and β_{12} have the following distributions

$$\hat{\beta}_{11} \sim N(\beta_{11}, \sigma^2/SSX_1) \text{ and } \hat{\beta}_{12} \sim N(\beta_{12}, \sigma^2/SSX_2),$$

where $SSX_1 = \sum_{j=1}^{N_1}(X_{1j} - \bar{X}_1)^2$ and $SSX_2 = \sum_{k=1}^{N_2}(X_{2k} - \bar{X}_2)^2$, $\bar{X}_1$ and $\bar{X}_2$ are the respective sample means of the X_{1j} and X_{2k} observations. Accordingly, $\hat{\beta}_{1D} = \hat{\beta}_{11} - \hat{\beta}_{12} \sim N\{\beta_{1D}, \sigma^2(1/SSX_1 + 1/SSX_2)\}$. On the other hand, $\hat{\sigma}^2 = SSE/\nu$ is the usual unbiased estimator of σ^2 where SSE is the error sum of squares and $\nu = N - 4$. Moreover, $SSE/\sigma^2 \sim \chi^2(\nu)$, where $\chi^2(\nu)$ are chi-square distribution with ν degrees of freedom. To detect the difference between two slope

coefficients in terms of H_0: $\beta_{11} = \beta_{12}$ versus H_1: $\beta_{11} \neq \beta_{12}$, the test statistic has the form

$$T = \frac{\hat{\beta}_{1D}}{\{\hat{\sigma}^2(1/SSX_1 + 1/SSX_2)\}^{1/2}}. \quad (4)$$

Under the null hypothesis H_0: $\beta_{11} = \beta_{12}$, the statistic has the distribution

$$T \sim t(v), \quad (5)$$

where $t(v)$ is a t distribution with degrees of freedom v. The null hypothesis is rejected at the significance level α if

$$|T| > t_{v,\alpha/2}, \quad (6)$$

where $t_{v,\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the distribution $t(v)$. Note that the inference setting is discussed here only from the perspective of a two-sided test. The same concepts may be readily extended to one-sided situations.

The statistical inferences about the heterogeneous slope effect are based on the conditional distribution of the continuous covariates. Therefore, the corresponding results would be specific to the particular values of the covariates. However, before conducting a research study, the actual values of covariates cannot be known in advance just as the primary responses. Under such circumstances, it is more suitable to employ the random or unconditional setup as explicated in Sampson [19]. The underlying similarities and differences between fixed and random models have also been thoroughly illuminated in Cramer and Appelbaum [20] and Raudenbush [21]. Despite the complexity associated with the unconditional properties of the test procedure, the tests of hypotheses and estimates of parameters remain the same under both conditional and unconditional frameworks. Hence, the usual rejection rule and critical value remain unchanged. The distinction between the two modeling approaches becomes important only when power and sample size calculations are to be made. Thus, it is vital to recognize the stochastic nature of the covariate variables and to evaluate the distribution of the test statistic over possible values of the covariates. In order to elucidate the critical notion of accommodating the distributional properties of the covariate variables, the continuous covariate variables $\{X_{1j}, j = 1,\ldots, N_1\}$ and $\{X_{2k}, k = 1,\ldots, N_2\}$ are assumed to have the independent normal distributions $N(\theta_1, \tau_1^2)$ and $N(\theta_2, \tau_2^2)$, respectively. It should be noted that the normality setting is commonly employed to provide a convenient framework for analytical derivation and theoretical discussion in interaction studies, for example, see Harwell [2], McClelland and Judd [22], O'Connor [23], and Shieh [14].

To help justify the contribution of current investigation, a brief review of the simple interaction model with two continuous covariates is presented here:

$$Y_i = \beta_I + X_i\beta_X + Z_i\beta_Z + X_iZ_i\beta_{XZ} + \xi_i, \quad (7)$$

where Y_i is the value of the response variable Y, X_i and Z_i are the known constants of the continuous covariates X and Z, ξ_i are *iid* $N(0,\omega^2)$ random errors for $i = 1,\ldots, N$, and β_I, β_X, β_Z, and β_{XZ} are unknown parameters. For the purpose of detecting the interaction effect in terms of the hypotheses H_0: $\beta_{XZ} = 0$ versus H_1: $\beta_{XZ} \neq 0$, it is important to examine the distributional property for the least squares estimator $\hat{\beta}_{XZ}$ of β_{XZ}:

$$\hat{\beta}_{XZ} \sim N(\beta_{XZ}, V(\hat{\beta}_{XZ})), \quad (8)$$

where $V(\hat{\beta}_{XZ}) = \omega^2 M$, M is the (3, 3) element of $(\mathbf{X}_C^{\mathrm{T}}\mathbf{X}_C)^{-1}$, where $\mathbf{X}_C = [\mathbf{X}_1 - \bar{\mathbf{X}}, \ldots, \mathbf{X}_N - \bar{\mathbf{X}}]^{\mathrm{T}}$, $\bar{\mathbf{X}} = \sum_{i=1}^{N}\mathbf{X}_i/N$, and $\mathbf{X}_i = [X_i, Z_i, X_iZ_i]^{\mathrm{T}}$ is the 3×1 column vector for values of covariates X_i,

Z_i, and their cross product X_iZ_i for $i = 1,\ldots, N$. The corresponding test statistic T_{XZ} is of the form

$$T_{XZ} = \frac{\hat{\beta}_{XZ}}{\{\hat{\omega}^2 M\}^{1/2}} \tag{9}$$

where $\hat{\omega}^2$ is the usual unbiased estimator of ω^2. When the null hypothesis H_0: $\beta_{XZ} = 0$ is true, the statistic T_{XZ} is distributed as $t(v)$, and H_0 is rejected at the significance level α if $|T_{XZ}| > t_{v,\alpha/2}$. At first sight, all of the model structure, tested hypothesis, and decision rule are similar to the prescribed results given in Eqs 4–6 for detecting the treatment by covariate interaction. However, the two test statistics T_{XZ} and T have different forms and distribution properties under alternative hypothesis. Specifically, an alternative expression for the centered design matrix $\mathbf{X}_C$ is $\mathbf{X}_C = [\mathbf{x}_C, \mathbf{z}_C, \mathbf{w}_C]$ where $\mathbf{x}_C$, $\mathbf{z}_C$, and $\mathbf{w}_C$ are the three $N \times 1$ column vectors of $\mathbf{X}_C$. Then, it can be shown that $M = (\mathbf{w}_C^{\mathrm{T}}\mathbf{M}_{AC}\mathbf{w}_C)^{-1}$, $\mathbf{M}_{AC} = \mathbf{I}_N - \mathbf{X}_{AC}(\mathbf{X}_{AC}^{\mathrm{T}}\mathbf{X}_{AC})^{-1}\mathbf{X}_{AC}^{\mathrm{T}}$ and $\mathbf{X}_{AC} = [\mathbf{x}_C, \mathbf{z}_C]$. The complex expression of M generally does not have a simple analytic distribution even though the two covariate variables X and Z may have a bivariate normal distribution. It should be obvious that the product XZ of two normally distributed variables does not have a normal distribution. Hence, it is inaccessible to obtain a transparent nonnull distribution for the test statistic T_{XZ} under random or unconditional framework with a given joint distribution of X and Z. Instead, Shieh [14] adopted a large-sample viewpoint and considered the asymptotic distribution of M. The resulting nonnull distribution and associated power function of the statistic T_{XZ} are considerably more complicated than the explications presented later for the T test of treatment by covariate interactions. Consequently, the power and sample size calculations of Shieh [14] for detecting interactions between two continuous variables in multiple regression analysis are not applicable for assessing interactions between grouping and continuous variables within the context of ANCOVA. In the following, particular attention is given to develop useful and specialized statistical techniques for power and sample size computations in assessing the difference between two regression slopes.

In general, the statistic T has the nonnull distribution for the given values of SSX_1 and SSX_2:

$$T|[SSX_1, SSX_2] \sim t(v, \Delta), \tag{10}$$

where $t(v,\Delta)$ is a noncentral t distribution with degrees of freedom v and noncentrality parameter

$$\Delta = \frac{\delta}{(1/SSX_1 + 1/SSX_2)^{1/2}}, \tag{11}$$

where $\delta = \beta_{1D}/\sigma$. It follows from Johnson, Kotz, and Balakrishnan [24] that the first moment of a noncentral t distribution is $E[T] = (v/2)^{1/2}\Gamma\{(v-1)/2\}\Delta/\Gamma\{v/2\}$, where $\Gamma\{\cdot\}$ is the gamma function. Hence, an unbiased estimator of the effect size δ is

$$\hat{\delta}_{UE} = \frac{(1/SSX_1 + 1/SSX_2)^{1/2}\Gamma\{v/2\}}{(v/2)^{1/2}\Gamma\{(v-1)/2\}} \cdot T = \frac{\Gamma\{v/2\}}{(v/2)^{1/2}\Gamma\{(v-1)/2\}} \cdot \frac{\hat{\beta}_{1D}}{\hat{\sigma}}.$$

To derive the nonnull distribution of T, an exact and sophisticated approach is to utilize the full distribution associated with SSX_1 and SSX_2. With the prescribed normal covariate assumptions, it can be readily established that $K_1 = SSX_1/\tau_1^2 \sim \chi^2(\kappa_1)$ and $K_2 = SSX_2/\tau_2^2 \sim \chi^2(\kappa_2)$ where $\kappa_1 = N_1-1$ and $\kappa_2 = N_2-1$. For ease of illustration, the two random variables of K_1 and K_2 are transformed to obtain $K = K_1 + K_2 \sim \chi^2(\kappa)$ and $B = K_1/K \sim \mathrm{Beta}\{\kappa_1/2, \kappa_2/2\}$ where $\mathrm{Beta}\{a, b\}$ is a beta distribution with degrees of freedom a and b. Note that the random variables K

and B are independent. Under the prescribed stochastic considerations of SSX_1 and SSX_2 in terms of K and B, the T statistic has the following two-stage distribution

$$T|[K,\ B] \sim t(v, \Delta_{KB}),\ K \sim \chi^2(\kappa),\ \text{and}\ B \sim \text{Beta}\{\kappa_1/2,\ \kappa_2/2\}. \tag{12}$$

where

$$\Delta_{KB} = \frac{\delta}{\{[1/(B_1\tau_1^2) + 1/(B_2\tau_2^2)]/K\}^{1/2}},$$

$B_1 = B$, and $B_2 = (1 - B)$. Hence, the resulting power function for comparing nonparallel regression lines is

$$\Psi_{KB}(\beta_{1D}) = E_K E_B[P\{|t(v, \Delta_{KB})| > t_{v,\alpha/2}\}], \tag{13}$$

where the expectation $E_K[\cdot]$ and $E_B[\cdot]$ is taken with respect to the distribution of K and B, respectively.

Alternatively, a simple and naive method to obtain a unconditional distribution of T is to substitute the two sum of squares SSX_1 and SSX_2 in Δ with the corresponding expected values $E[SSX_1] = \kappa_1\tau_1^2$ and $E[SSX_2] = \kappa_2\tau_2^2$. Consequently, the distribution of T can be approximated by a noncentral t distribution as

$$T \dot{\sim} t(v, \Delta_A), \tag{14}$$

where

$$\Delta_A = \frac{\delta}{\{[1/(b_1\tau_1^2) + 1/(b_2\tau_2^2)]/\kappa\}^{1/2}},$$

$b_1 = \kappa_1/\kappa$, $b_2 = \kappa_1/\kappa$, and $\kappa = \kappa_1 + \kappa_2$. The corresponding power function for the test for heterogeneity of regression slopes can be expressed as

$$\Psi_A(\beta_{1D}) = P\{|t(v, \Delta_A)| > t_{v,\alpha/2}\}. \tag{15}$$

On the other hand, Dupont and Plummer [13] presented a relatively more simplified power function for the test of difference between two regression slopes:

$$\Psi_{DP}(\beta_{1D}) = P\{t(v) < \Delta_{DP} - t_{v,\alpha/2}\} + P\{t(v) < -\Delta_{DP} - t_{v,\alpha/2}\}. \tag{16}$$

where

$$\Delta_{DP} = \frac{\delta}{\{[1/(p_1\tau_1^2) + 1/(p_2\tau_2^2)]/N\}^{1/2}},$$

$p_1 = N_1/N$ and $p_2 = 1 - p_1$. Although the two noncentrality parameters Δ_A and Δ_{DP} are quite similar, especially when the sample size N is large, the two approximate power functions Ψ_A and Ψ_{DP} have a crucial difference. Note that the power function Ψ_A involves a noncentral t distribution $t(v,\Delta_A)$, whereas Ψ_{DP} is formulated through a shifted t distribution $t(v) + \Delta_{DP}$). It is well known that if $Z \sim N(0, 1)$ then $X = (Z + \mu) \sim N(\mu, 1)$ where μ is a constant. However, the result does not generalize to the case of t distribution, i.e., if $t \sim t(df)$ then $Y = (t + \mu)$ does not follow a noncentral t distribution $t(df, \mu)$ with noncentrality parameter μ and degrees of freedom df. A random variable Y is said to have a noncentral t distribution $t(df, \mu)$ if and only if $Y = (Z + \mu)/(W/df)^{1/2}$ where $Z \sim N(0, 1)$, $W \sim \chi^2(df)$, and Z and W are independent. Essentially, Dupont and Plummer [13] extended the results under normal theory in Dupont and Plummer [25] to the case of noncentral t distributions in the comparison of two regression

slopes. The resulting formulation suffers the absence of proper theoretical justification. Despite the computational appeal of the approximate power function Ψ_{DP}, the prescribed analytic issue induces a fundamental question about its general adequacy as a reliable procedure.

It is essential to note that all the power functions Ψ_{DP}, Ψ_A and Ψ_{KB} depend on the difference between two coefficients $\{\beta_{11}, \beta_{12}\}$ and error variance σ^2 through the standardized effect δ. Under the prescribed stochastic assumptions for the covariate variables, these power functions rely on the covariate variances $\{\tau_1^2, \tau_2^2\}$ through the associated noncentrality parameter, but not the mean values of covariate variables $\{\theta_1, \theta_2\}$. Moreover, the approximate formulations of Ψ_{DP} and Ψ_A only involve the central t and noncentral t distributions, whereas the normal covariate distributions lead to the unique and more complex conditional property of Ψ_{KB} on the chi-square distribution and beta distribution. It can be shown that the noncentrality terms Δ_{DP}, Δ_A, and Δ_{KB} are asymptotically equivalent as sample size goes to infinity. Therefore, the three power functions Ψ_{DP}, Ψ_A, and Ψ_{KB} have the same large sample properties. Despite the close resemblance between the three power formulas, the corresponding behaviors for finite sample obviously differ. Their relative performance of power calculations will be appraised in the numerical investigations.

For planning research design, the power formulas can be employed to determine the sample sizes N_1 and N_2 needed to attain the specified power $(1 - \beta)$ through a simple iterative search for the chosen significance level α and parameter settings. In practice, a research study requires adequate statistical power and sufficient sample size to detect scientifically credible effects. It is sensible that the corresponding power calculations and sample size determinations must be considered in the planning stage of a study. Consequently, it is of theoretical importance to evaluate the potential discrepancy between the three procedures in power and sample size calculations. In view of the wide variety of practical situations, the presumed normal covariate distribution merely provides a convenient and important situation. Evidently, the degree of robustness to nonnormal covariates for the resulting power and sample size procedures is also an essential issue and requires further sensitivity assessments.

Simulation study

To justify the distinct advantage of the suggested exact approach and the potential deficiency of the approximate methods, numerical examinations of power and sample size calculations were conducted in two studies under a wide variety of model configurations. The first investigation focuses on the situations with normal covariate variables, whereas several notable scenarios of non-normal covariates are examined in the subsequent appraisal.

Study I

For the purpose of explicating the critical discrepancy between the three power functions Ψ_{DP}, Ψ_A, and Ψ_{KB} in using covariate information, the two covariates X_1 and X_2 are assumed to have normal distributions with variances $\{\tau_1^2, \tau_2^2\} = \{1, 1\}$ and $\{1, 3\}$ for balanced design with $N_1 = N_2$ and $\{\tau_1^2, \tau_2^2\} = \{1, 1\}$, $\{1, 3\}$, and $\{3, 1\}$ for unbalanced design with $N_2 = 3N_1$. As noted earlier, the power functions do not depend on the covariate means θ_1 and θ_2. Without loss of generality, they are set as $\theta_1 = \theta_2 = 0$. In addition, the selected configurations of treatment means and error variance are $\beta_{11} = 0.50$ and 0.75, $\beta_{12} = 0$, and $\sigma^2 = 1$. Hence, the resulting standardized effect size has two different values $\delta = 0.50$ and 0.75. Overall these considerations result in a total of 10 different combined arrangements. These combinations of different covariate structures, effect magnitudes, and sample size allocations were chosen to represent as much as possible the extent of characteristics that are likely to be encountered in actual applications.

Table 1. Computed sample size, estimated power, and simulated power when $\delta = 0.50$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$.

	Dupont and Plummer (1998)				Approximate method				Exact approach			
Covariate variance	Sample sizes	Estimated power	Simulated power	Error	Sample sizes	Estimated power	Simulated power	Error	Sample sizes	Estimated power	Simulated power	Error
{1, 1}	{64, 64}	0.8013	0.7773	0.0240	{65, 65}	0.8015	0.7859	0.0156	{67, 67}	0.8026	0.8000	0.0026
{1, 3}	{43, 43}	0.8011	0.7804	0.0207	{44, 44}	0.8015	0.7844	0.0171	{46, 46}	0.8037	0.8065	−0.0028
{1, 1}	{43, 129}	0.8059	0.7894	0.0165	{44, 132}	0.8076	0.7977	0.0099	{45, 135}	0.8033	0.8014	0.0019
{1, 3}	{36, 108}	0.8068	0.7831	0.0237	{37, 111}	0.8007	0.7943	0.0134	{38, 114}	0.8015	0.8011	0.0004
{3, 1}	{22, 66}	0.8103	0.7674	0.0429	{23, 69}	0.8165	0.7920	0.0245	{24, 72}	0.8122	0.8169	−0.0047

https://doi.org/10.1371/journal.pone.0177682.t001

With the prescribed specifications, the required sample sizes were computed for the three procedures with the chosen power value and significance level. Throughout this empirical investigation, the significance level and nominal power are fixed as $\alpha = 0.05$ and $1 - \beta = 0.80$, respectively. The computed sample sizes associate with the effect size $\delta = 0.50$ and 0.75 are presented in Tables 1 and 2, respectively. For ease of illustration, the total sample sizes of the exact approach for $\delta = 0.50$ and 0.75 are plotted in Figs 1 and 2, respectively.

The graphs show that, for fixed values of sample size ratio r and covariate variance τ_1^2, the total sample sizes N decrease with increasing covariance variance τ_2^2. It is clear that the computed sample sizes in Table 1 are larger than those in Table 2 when all other characteristics are the same. More importantly, the results show that the calculated sample sizes of the exact approach differ from those of the two approximate procedures for all ten cases. The sample sizes of the approximate methods are relatively smaller than those of the exact approach. Also, the discrepancy are slightly larger for $\delta = 0.75$ in Table 2 than those of $\delta = 0.50$ in Table 1. In order to evaluate the accuracy of the power functions, the estimated power or computed power are also listed. Because of the underlying metric of integer sample sizes, the attained values are marginally larger than the nominal level for all three procedures.

Then, Monte Carlo simulation studies were performed to evaluate the accuracy of the sample size calculations. With the computed sample sizes, parameter configurations, and nominal power, estimates of the true power were computed via Monte Carlo simulation of 10,000 independent data sets. For each replicate, N_1 and N_2 covariate values were generated from the selected normal distributions. The resulting values of covariate variables in turn determined the mean responses for generating N_1 and N_2 normal outcomes with the designated ANCOVA

Table 2. Computed sample size, estimated power, and simulated power when $\delta = 0.75$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$.

	Dupont and Plummer (1998)				Approximate method				Exact approach			
Covariate variance	Sample sizes	Estimated power	Simulated power	Error	Sample sizes	Estimated power	Simulated power	Error	Sample sizes	Estimated power	Simulated power	Error
{1, 1}	{29, 29}	0.8008	0.7682	0.0326	{30, 30}	0.8014	0.7745	0.0269	{32, 32}	0.8045	0.8072	−0.0027
{1, 3}	{20, 20}	0.8068	0.7467	0.0601	{21, 21}	0.8080	0.7715	0.0365	{23, 23}	0.8135	0.8156	−0.0021
{1, 1}	{20, 60}	0.8180	0.7687	0.0493	{20, 60}	0.8016	0.7719	0.0297	{22, 66}	0.8125	0.8145	−0.0020
{1, 3}	{17, 51}	0.8236	0.7710	0.0526	{17, 51}	0.8020	0.7736	0.0284	{19, 57}	0.8124	0.8169	−0.0045
{3, 1}	{10, 30}	0.8068	0.7194	0.0874	{11, 33}	0.8211	0.7713	0.0498	{12, 36}	0.8126	0.8181	−0.0055

https://doi.org/10.1371/journal.pone.0177682.t002

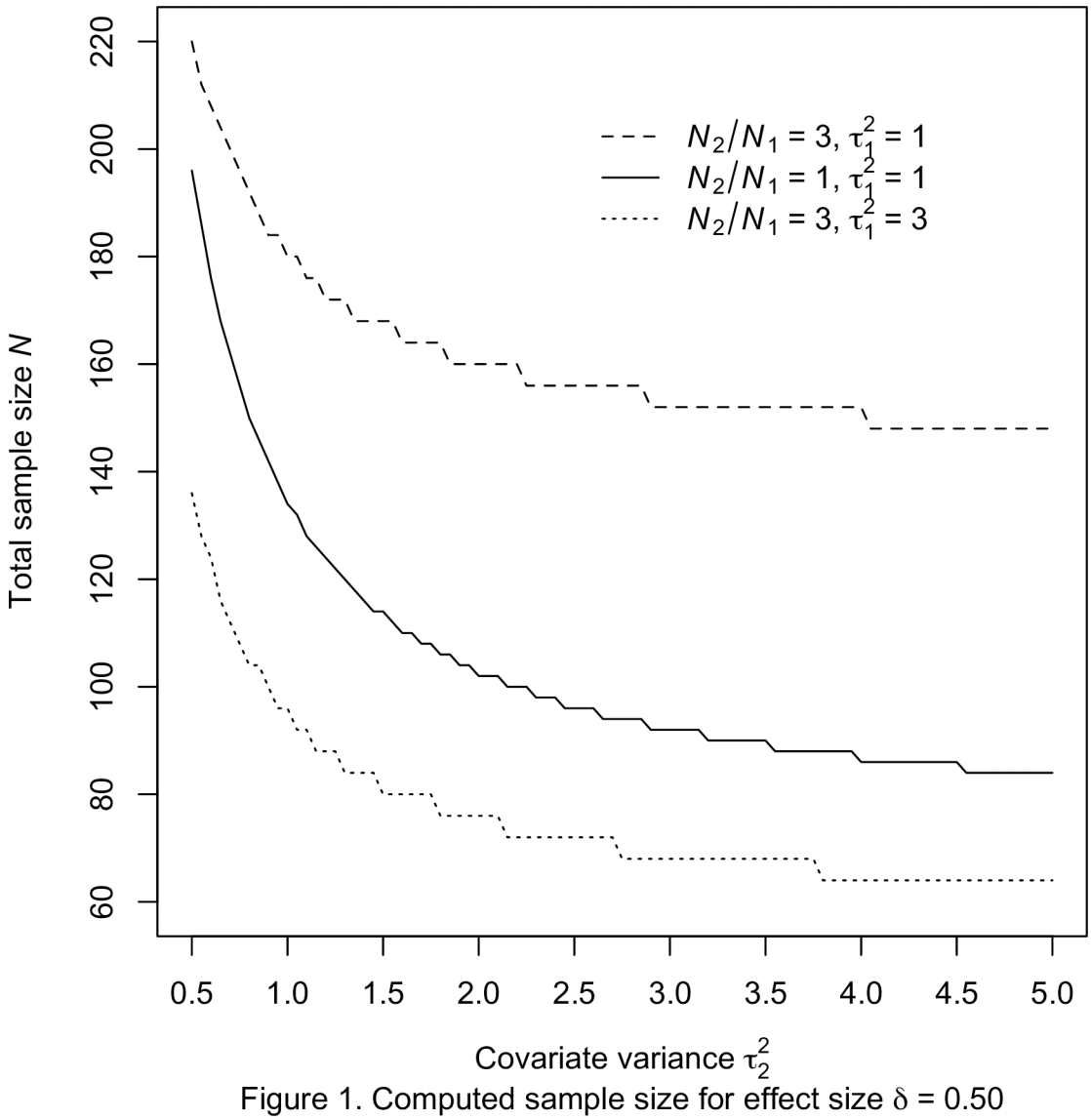

Fig 1. Computed sample size for effect size $\delta = 0.50$.

https://doi.org/10.1371/journal.pone.0177682.g001

designs. Next, the test statistic T was computed and the simulated power was the proportion of the 10,000 replicates whose test statistics $|T|$ exceeded the corresponding critical value $t_{v,0.025}$. Therefore, the adequacy of the approximate and exact sample size procedures is determined by the error (= estimate power–simulated power) between the estimated power computed from analytic formulas and the simulated power of Monte Carlo study. The simulated power and error are also summarized in Tables 1 and 2 for all 10 design schemes.

It is noticeable from the results that there exists a close agreement between the estimated power and the simulated power for the proposed exact sample size procedure regardless of the model configurations. Specifically, all the incurred errors of the 10 designs are all within the small range of –0.0055 to 0.0026. In contrast, the estimated powers for the two approximate methods are consistently larger than the simulated powers for all 10 settings in Tables 1 and 2. In particular, the errors associated with Dupont and Plummer's [13] procedure are {0.0240,

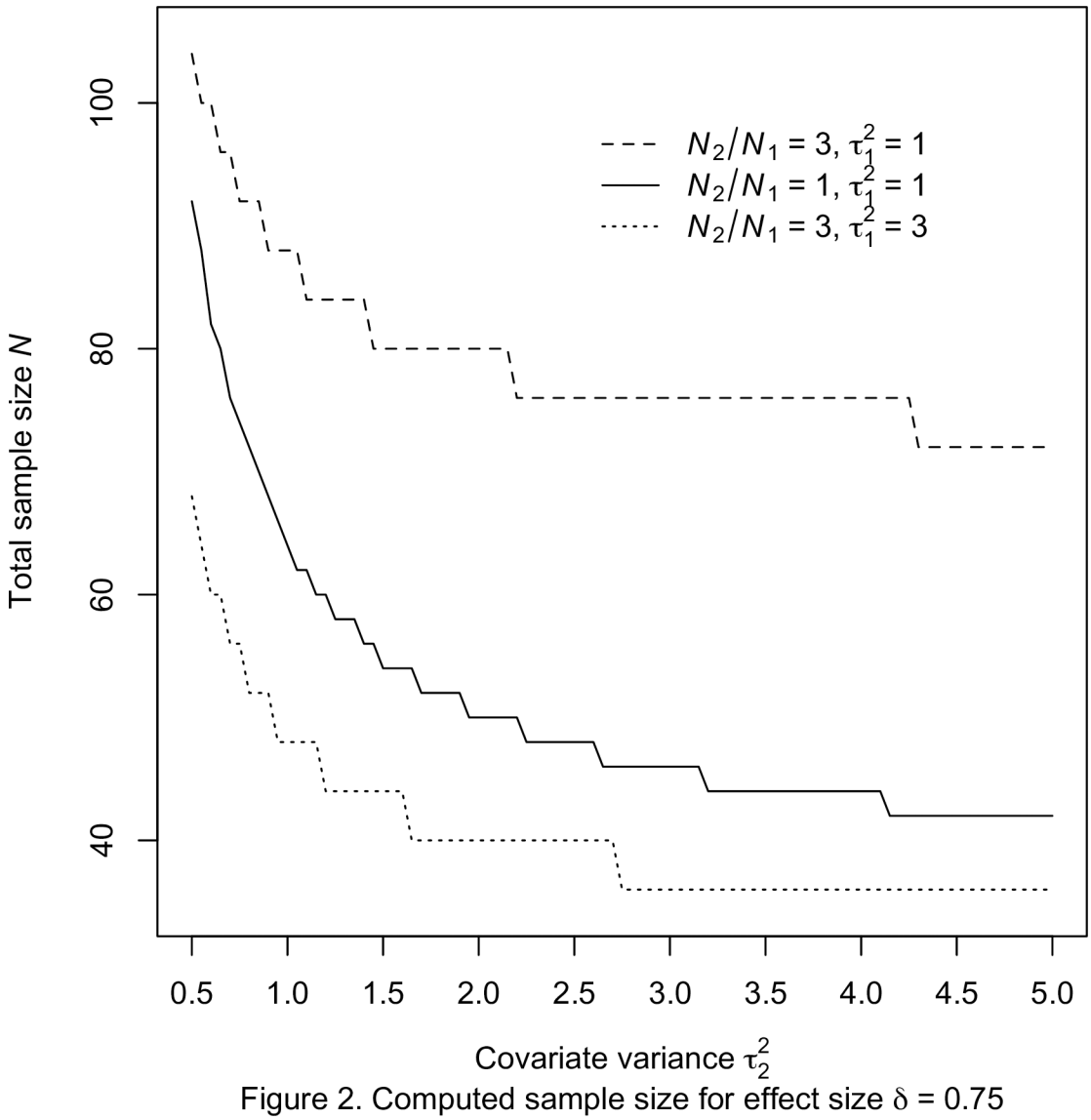

Fig 2. Computed sample size for effect size $\delta = 0.75$.

https://doi.org/10.1371/journal.pone.0177682.g002

0.0207, 0.0165, 0.0237, 0.0429} and {0.0326, 0.0601, 0.0493, 0.0526, 0.0874} for $\delta = 0.50$ and 0.75 in Tables 1 and 2, respectively. For the approximate method with power function Ψ_A, the corresponding errors of the ten cases in Tables 1 and 2 are {0.0156, 0.0171, 0.0099, 0.0134, 0.0245} and {0.0269, 0.0365, 0.0297, 0.0284, 0.0498} for $\delta = 0.50$ and 0.75, respectively. Although some of the differences are not substantial, it delineates a clear pattern that the accuracy of the approximate power functions deteriorates to some degree for smaller sample sizes, especially for the simple method of Dupont and Plummer [13]. Furthermore, the magnitudes of errors correspond to the direct-paring cases (when larger covariate variance is paired with larger sample size) are relative smaller than those of the inverse-pairing situations (when larger covariate variance is paired with smaller sample size). Note that the resulting errors of Dupont and Plummer's [13] procedure associated with $\{\tau_1^2, \tau_2^2\} = \{1, 3\}$ and $\{N_1, N_2\} = \{36, 108\}$ and {17, 51} under direct-pairing are 0.0237 and 0.0526 in Tables 1 and 2, respectively. However,

the counterparts of inverse-pairing setting with $\{\tau_1^2, \tau_2^2\} = \{3, 1\}$ and $\{N_1, N_2\} = \{22, 66\}$ and $\{10, 30\}$ are much larger with 0.0429 and 0.0874 for $\delta = 0.50$ and 0.75, respectively. These realizations imply that the magnitude of sample sizes plays an essential role in the performance of the approximate methods. More importantly, the adequacy of the approximate power formulas and sample size procedures varies with model configurations. In contrast, the numerical performance suggests that the exact methodology performs fairly well for the range of model specifications considered here.

Study II

The described exact power function is obtained under the essential framework that the covariate variables have normal distributions. Instead of using the full features, the approximate power formula Ψ_A only relies on the partial information of second moments or variances of the covariates. At first sight, the simplified method may be more robust than the exact approach to the violation of normality assumption of the covariates. To further illuminate the sensitivity issues and profound implications of the two distinct techniques, power and sample size calculations were also conducted for the scenarios with non-normal covariates. Due to the undesired and inferior performance of Dupont and Plummer's [13] technique, their method is not considered in this examination.

Specifically, the two covariates are assumed to have five different sets of distributions: Beta, Exponential, Gamma, Laplace, and Uniform. For ease of comparison, the designated distributions were constructed to have variances $\{\tau_1^2, \tau_2^2\} = \{1, 1\}$ and $\{1, 3\}$. Moreover, only balanced designs were considered and the treatment means and error variance were fixed as $\beta_{11} = 0.50$, $\beta_{12} = 0$, and $\sigma^2 = 1$. Hence, the required sample sizes and estimated powers associated with the exact procedure remain identical for the five different distributions. Unlike the previous study, the estimated powers and related evaluations of the approximate method were computed with the sample sizes determined by the exact approach. Table 3 summarizes the empirical results of the ten combined structures of covariate distribution and associated variance. In the case of Beta distribution, the actual two pairs of Beta covariates are $X_1 \sim \text{Beta}(2, 5)/c_1$ and $X_2 \sim \text{Beta}(2, 5)/c_1$, and $X_1 \sim \text{Beta}(2, 5)/c_1$ and $X_2 \sim \text{Beta}(2, 5)/c_2$ where c_1 and c_2 are selected such that the resulting variances are 1 and 3, respectively. On the other hand, the parameter specifications of

Table 3. Computed sample size, estimated power, and simulated power when $\delta = 0.50$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$.

		Approximate method			Exact approach		
Covariate distributions	Sample sizes	Estimated power	Simulated power	Error	Estimated power	Simulated power	Error
Beta(2, 5)* and Beta(2, 5)*	{67, 67}	0.8135	0.7973	0.0162	0.8026	0.7973	0.0053
Beta(2, 5)* and Beta(2, 5)**	{46, 46}	0.8194	0.7960	0.0234	0.8037	0.7960	0.0077
Exponential(1) and Exponential(1)	{67, 67}	0.8135	0.7775	0.0360	0.8026	0.7775	0.0251
Exponential(1) and Exponential($3^{1/2}$)	{46, 46}	0.8194	0.7697	0.0497	0.8037	0.7697	0.0340
Gamma(2, $1/2^{1/2}$) and Gamma(2, $1/2^{1/2}$)	{67, 67}	0.8135	0.7905	0.0230	0.8026	0.7905	0.0121
Gamma(2, $1/2^{1/2}$) and Gamma(2, $(3/2)^{1/2}$)	{46, 46}	0.8194	0.7830	0.0364	0.8037	0.7830	0.0207
Laplace($2^{1/2}$) and Laplace($2^{1/2}$)	{67, 67}	0.8135	0.7927	0.0208	0.8026	0.7927	0.0099
Laplace($2^{1/2}$) and Laplace($(2/3)^{1/2}$)	{46, 46}	0.8194	0.7814	0.0380	0.8037	0.7814	0.0223
Uniform(–1/2, 1/2) and Uniform(–1/2, 1/2)	{67, 67}	0.8135	0.8115	0.0020	0.8026	0.8115	–0.0089
Uniform(–1/2, 1/2) and Uniform(–3, 3)	{46, 46}	0.8194	0.8095	0.0099	0.8037	0.8095	–0.0058

*Beta(2, 5) is scaled to have a variance 1

**Beta(2, 5) is scaled to have a variance 3.

https://doi.org/10.1371/journal.pone.0177682.t003

the other four types of distribution can be found in Table 3. Similar to the numerical assessments in Study I, Table 3 presents the computed sample sizes, estimated powers, simulated powers, and associated errors of the two competing procedures.

A detailed inspection of the findings in Table 3 reveals that the performance of both the contending procedures is affected by the non-normal covariate settings, especially for the Exponential cases. However, it is important to note that the approximate technique incurs larger estimated powers and errors between estimated power and simulated power than the exact approach. The only exceptions occurred with the Uniform covariate distribution that the exact procedure does not have a clear advantage over the approximate method. Conceivably, the degree of robustness of the suggested exact technique presumably depends on the extent of how badly covariate distributions deviate from normality assumption. Nonetheless, these empirical evidences show that the exact procedure give acceptable results even for the non-normal covariates. In view of the potentially diverse treatment and covariate configurations of ANCOVA studies, it appears that the exact approach is relatively more consistent and accurate than the approximate method to be considered as a general tool.

Results

The implementation of the suggested power and sample size calculations involves specialized programs not currently available in prevailing statistical packages. To exemplify the computational aspects of the developed algorithms for design planning, the numerical demonstration of evaluating two treatments for gingivitis in Fleiss [3, Section 7.3] is reexamined here. The data consists of measurements of patients before and after treatment on a modification of the Loe and Silness [26] index of gingivitis. A higher value indicates a more severe level of gingivitis. Accordingly, the response variable of ANCOVA is the post-treatment measurement with the pretreatment value serving as the covariate. It should be note that the illustration in Fleiss [3] does not address the power and sample size issues. Moreover, the emphasis of this numerical demonstration is on the typical research scenario most frequently encountered in the planning stage of an ANCOVA study.

Due to the prospective nature of advance research planning, the general guidelines suggest that typical sources like published finding or expert opinion can offer plausible and reasonable planning values for the model characteristics, such as treatment effects, variance component, and covariate properties. To explicate the essential processes, the prescribed data of comparing two treatments of gingivitis is employed to provide planning values of the model parameters and covariate configurations for related gingivitis studies. Specifically, the summary statistics yield the designated treatment effects and variance component: $\beta_{11} = 0.8502$, $\beta_{12} = 0.4008$, and $\sigma^2 = 0.04$. In addition, the covariate variances are obtained from the reported pretreatment values as $\tau_1^2 = 0.0646$ and $\tau_2^2 = 0.0526$. With the sample sizes of $\{N_1, N_2\} = \{74, 64\}$ and significance level $\alpha = 0.05$, the achieved power can be readily computed with the supplemental programs (Programs A and C). The result shows that the achieved power of the particular unbalanced design is $\Psi_{KB} = 0.8650$ which falls between the two fairly common levels of 0.80 and 0.90. Therefore, the power calculation suggests that the designated configurations warrant a decent chance of detecting the slope difference between two treatment groups.

Alternatively, under the notion of a balanced design, the presented algorithms (Programs B and D) reveal that the equal sample sizes of $\{N_1, N_2\} = \{69, 69\}$ yield the power of 0.8694. It is interesting to note that, although the two sample size schemes {74, 64} and {69, 69} have the identical total sample size 138, the balanced design has a slightly advantage over the unbalanced structure in power performance. For an illustration of sample size determination for planning balanced study, detailed computations show that the balanced sample sizes of $\{N_1,$

N_2} = {58, 58} and {77, 77} are needed to achieve the target powers of 0.80 and 0.90, respectively. It is noted above, because of the sample sizes need to be integer values in practice, that the attained power is marginally greater than the nominal power level. Here, the corresponding actual powers of the two sample size designs are 0.8043 and 0.9038, respectively. These vital configurations are incorporated in the user specifications of the SAS/IML [13] and R [14] programs presented in the supplemental files. With the prescribed explications, users can easily identify the statements containing the exemplifying values in the computer code and then modify the program to accommodate their own model specifications.

Conclusions and discussion

Within the context of ANCOVA, an underlying assumption is the parallelism of the regression lines associating the criterion variable with the covariate. It has been emphasized that the homogeneity of covariate regression slopes is the most important statistical assumption in ANCOVA. However, there are theoretical reasons and empirical evidences to document non-parallel phenomenon of regression lines across many scientific fields. Although the test of the hypothesis of parallel regression lines is a simple and straightforward procedure, the corresponding analytic derivations and computational algorithms of power and sample size determinations have not been examined in the literature. Conceivably, the corresponding power analysis and sample size determination must also be considered before it can be adopted as a general methodology in practice. To facilitate proper use and implication of traditional ANCOVA and extended alternatives, this article presents both pedagogical explication and numerical appraisal of power and sample size procedures for the detection of heterogeneity between two covariate regression coefficients. Despite the simplicity, this scenario embodies all the essential notion and critical feature of ANCOVA that can be useful in undertaking similar considerations for the more involved multi-group situations.

The existing method of Dupont and Plummer (1998) seems to provide a simple solution and maintains reasonable accuracy for some model configurations. However, no research to date has properly examined its properties both analytically and empirically. The presented analytic explication and empirical results showed that the approximate formula of Dupont and Plummer [13] does not guarantee to give accurate power and sample size calculations. The proposed exact approach has the distinct feature of accommodating the full distributional properties of normal covariates whereas the simplified approximate methods only utilize the partial information of covariate variances. It is important to note that although Glueck and Muller [27] and Shieh [28] considered the problem of adjusting power for random covariates in multivariate linear models, their model formulations do not cover the interaction effects between treatment groups and continuous covariates. Hence, the corresponding power and sample size procedures do not applied to the detection of slope heterogeneity considered here. Moreover, due to the complexity of multivariate settings, only moments of the covariate variables are employed in the power formulas presented in Glueck and Muller [27] and Shieh [28]. Consequently, their methods do not take into account the full distributional features of covariate variables. In view of the overall accuracy and robustness, the exact approach is recommended over the approximate methods as a reliable tool in practical applications. The supporting SAS/IML [15] and R [16] computer algorithms will yield accurate power calculations and sample size determinations provided that all the required information is properly specified.

Supporting information

S1 File. SAS programs.
(DOCX)

S2 File. R programs.
(DOCX)

Author Contributions

Conceptualization: GS.

Data curation: GS.

Formal analysis: GS.

Funding acquisition: GS.

Investigation: GS.

Methodology: GS.

Project administration: GS.

Resources: GS.

Software: GS.

Supervision: GS.

Validation: GS.

Visualization: GS.

Writing – original draft: GS.

Writing – review & editing: GS.

References

1. Glass G. V, Peckham P. D., & Sanders J. R. (1972). Consequences of failure to meet assumptions underlying the analysis of variance and covariance. Review of Educational Research, 42, 237–288.
2. Harwell M. (2003). Summarizing Monte Carlo results in methodological research: The single-factor, fixed-effects ANCOVA case. Journal of Educational and Behavioral Statistics, 28, 45–70.
3. Fleiss J. L. (2011). Design and analysis of clinical experiments (Vol. 73). New York, NY: Wiley.
4. Huitema B. (2011). The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies (Vol. 608). New York, NY: Wiley.
5. Maxwell S. E., & Delaney H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.
6. Hauck W. W., Anderson S., & Marcus S. M. (1998). Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clinical Trials, 19, 249–256.
7. Hernandez A. V., Steyerberg E. W., & Habbema J. D. F. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. Journal of Clinical Epidemiology, 57, 454–460. https://doi.org/10.1016/j.jclinepi.2003.09.014 PMID: 15196615
8. Pocock S. J., Assmann S. E., Enos L. E., & Kasten L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparison in clinical trial reporting: Current practice and problems. Statistics in Medicine, 21, 2917–2930. https://doi.org/10.1002/sim.1296 PMID: 12325108
9. Raab G. M., Day S., & Sales J. (2000). How to select covariates to include in the analysis of a clinical trial. Controlled Clinical Trials, 21, 330–342. PMID: 10913808
10. Hauck W. W., Neuhaus J.M., Kalbfleisch J.D., & Anderson S. (1991). A consequence of omitted covariates when estimating odds ratios. Journal of Clinical Epidemiology, 44, 77–81. PMID: 1986061
11. Gail M. H., Wieand S., & Piantadosi S. (1984). Biased estimates of treatment effect in randomized experiments with non-linear regression and omitted covariates. Biometrika, 71, 431–44.

12. Negassa A., & Hanley J. A. (2007). The effect of omitted covariates on confidence interval and study power in binary outcome analysis: A simulation study. Contemporary Clinical Trials, 28, 242–248. https://doi.org/10.1016/j.cct.2006.08.007 PMID: 17011835
13. Dupont W. D., & Plummer W. D. (1998). Power and sample size calculations for studies involving linear regression. Controlled Clinical Trials, 19, 589–601. PMID: 9875838
14. Shieh G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables: Power and sample size considerations. Organizational Research Methods, 12, 510–528.
15. SAS Institute. SAS/IML User's Guide, Version 9.3. Cary, NC: SAS Institute Inc; 2014.
16. R Development Core Team. R: A language and environment for statistical computing [Computer software and manual]; 2016. Retrieved from http://www.r-project.org.
17. Rogosa D. (1980). Comparing nonparallel regression lines. Psychological Bulletin, 88, 307–321.
18. Rutherford A. (1992). Alternatives to traditional analysis of covariance. British Journal of Mathematical and Statistical Psychology, 45, 197–223.
19. Sampson A. R. (1974). A tale of two regressions. Journal of the American Statistical Association, 69, 682–689.
20. Cramer E. M., & Appelbaum M. I. (1978). The validity of polynomial regression in the random regression model. Review of Educational Research, 48, 511–515.
21. Raudenbush S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. Psychological Methods, 2, 173–185.
22. McClelland G. H., & Judd C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. Psychological Bulletin, 114, 376–390. PMID: 8416037
23. O'Connor B. P. (2006). Programs for problems created by continuous variable distributions in moderated multiple regression. Organizational Research Methods, 9, 554–567.
24. Johnson N. L., Kotz S., & Balakrishnan N. (1995). Continuous univariate distributions (2nd ed., Vol. 2). New York, NY: Wiley.
25. Dupont W. D., & Plummer W. D. (1990). Power and sample size calculations: A review and computer program. Controlled Clinical Trials, 11, 116–128. PMID: 2161310
26. Loe H., & Silness J. (1963). Periodontal disease in pregnancy. Acta Odontologica Scandinavica, 21, 533–551. PMID: 14121956
27. Glueck D. H., & Muller K.E. (2003). Adjusting power for a baseline covariate in linear models. Statistics in Medicine, 22, 2535–2551. https://doi.org/10.1002/sim.1341 PMID: 12898543
28. Shieh G. (2005). Power and sample size calculations for multivariate linear models with random explanatory variables. Psychometrika, 70, 347–358.