

Small-sample performance and underlying assumptions of a bootstrap-based inference method for a general analysis of covariance model with possibly heteroskedastic and nonnormal errors

Georg Zimmermann,^{1,2,3}  Markus Pauly⁴ and Arne C Bathke^{1,5}

Statistical Methods in Medical Research
2019, Vol. 28(12) 3808–3821

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0962280218817796

journals.sagepub.com/home/smm



Abstract

It is well known that the standard F test is severely affected by heteroskedasticity in unbalanced analysis of covariance models. Currently available potential remedies for such a scenario are based on heteroskedasticity-consistent covariance matrix estimation (HCCME). However, the HCCME approach tends to be liberal in small samples. Therefore, in the present paper, we propose a combination of HCCME and a wild bootstrap technique, with the aim of improving the small-sample performance. We precisely state a set of assumptions for the general analysis of covariance model and discuss their practical interpretation in detail, since this issue may have been somewhat neglected in applied research so far. We prove that these assumptions are sufficient to ensure the asymptotic validity of the combined HCCME-wild bootstrap analysis of covariance. The results of our simulation study indicate that our proposed test remedies the problems of the analysis of covariance F test and its heteroskedasticity-consistent alternatives in small to moderate sample size scenarios. Our test only requires very mild conditions, thus being applicable in a broad range of real-life settings, as illustrated by the detailed discussion of a dataset from preclinical research on spinal cord injury. Our proposed method is ready-to-use and allows for valid hypothesis testing in frequently encountered settings (e.g., comparing group means while adjusting for baseline measurements in a randomized controlled clinical trial).

Keywords

Heteroskedasticity-consistent covariance matrix estimator, analysis of covariance, rare disease, resampling, small sample, wild bootstrap

1 Introduction

1.1 The analysis of covariance model and its assumptions

Consider the frequently encountered situation where several groups of subjects are being compared with respect to a continuous outcome variable. For the statistical comparison of the group means, the analysis of variance (ANOVA) is often used. However, in many instances, it is reasonable to account for one or several covariates, such as baseline measurements or variables which are thought to be related to the outcome. A recently published EMA guideline for clinical trials recommends adjusting for any variable which is at least moderately associated

¹Department of Mathematics, Paris Lodron University, Salzburg, Austria

²Spinal Cord Injury and Tissue Regeneration Centre Salzburg, Paracelsus Medical University, Salzburg, Austria

³Department of Neurology, Christian Doppler Medical Centre and Centre for Cognitive Neuroscience, Salzburg, Austria

⁴Institute of Statistics, University of Ulm, Ulm, Germany

⁵Department of Statistics, University of Kentucky, Lexington, KY, USA

Corresponding author:

Georg Zimmermann, Department of Mathematics, Paris Lodron University, Hellbrunner Strasse 34, A 5020 Salzburg, Austria.

Email: georg.zimmermann@sbg.ac.at

with the primary outcome.¹ For this purpose, the analysis of covariance (ANCOVA) is an appropriate tool, which is used with the aim of increasing the inferential power, and reducing bias and variance of the effect estimators.² The ANCOVA has been applied in many research disciplines, ranging from studies about Alzheimer's disease³ to pharmaceutical issues,⁴ educational⁵ and fishery research,⁶ just to name a few.

There have been controversial discussions concerning the appropriate use and interpretation of ANCOVA in various settings.^{5,7–11} Apart from that, it is well known that, as an inference method, it also relies heavily on the assumptions of homoskedasticity and normality of the errors. It has been shown in simulation studies that the violation of one or more of these assumptions can seriously affect the ANCOVA F test in terms of maintenance of the prespecified type I error level and power.^{2,12}

Many of the proposed solutions to tackle this problem can be grouped into one of the following two approaches. Namely, on the one hand, some authors essentially kept the fully parametric ANCOVA model, but tried to derive test statistics which are more robust against violations of the above-mentioned assumptions. This approach had already been considered several decades ago¹³ and has recently been enriched by different bootstrap variations.¹⁴ On the other hand, leaving the parametric model in favour of a nonparametric approach has received attention. In this regard, both introductory papers explaining the proper application of ANCOVA to real-life data,^{15–17} as well as more theoretic contributions^{18–22} have been published. However, the latter approaches still impose some restrictions regarding either the number of groups or the number of covariates. Moreover, only moderate to large sample size scenarios have been considered so far, with a minimum number of 40 subjects per group.²¹ Thus, the small-sample performance of those methods remains unknown. However, group sizes below that level are encountered quite frequently and there is a need for precise statistical methods for such situations. Examples cover data from preclinical research as well as studies on rather rare diseases (e.g., spinal cord injury (SCI) or multiple sclerosis). In one single paper, the authors indeed examined the small-sample performance of the tests they proposed with respect to the maintenance of the nominal α level. They found that finite-sample properties also depended on the number of groups.¹⁸ Likewise, the enriched parametric ANCOVA approach lacks sufficient evidence regarding its performance in finite-sample situations. For example, in simulation studies, only the homoskedasticity assumption was relaxed, but the normality assumption was maintained.¹⁴

1.2 Heteroskedasticity-consistent covariance matrix estimation in regression analysis

Relaxing the model assumptions has been an important focus of research in the field of regression analysis for some decades, too. Especially with regard to heteroskedasticity of the errors, an important contribution was White's heteroskedasticity-consistent covariance matrix estimator (HCCME), and the derivation of the asymptotic distributions of its corresponding test statistics.²³ In medical studies, this approach has been used to some extent, too, both in applied branches such as diffusion tensor imaging²⁴ and genome-wide microarray analyses,²⁵ as well as in more methodological papers.^{26,27} However, the statement that HCCMEs are hardly known outside the statistical audience²⁸ still appears to be a fair assessment. Moreover, the methods discussed so far either lack sufficient generality, maintaining strong assumptions like the normality of the errors,²⁶ or they do not consider the assumptions underlying HCCMEs at all.^{25–28} In addition to that, results of simulation studies conducted in the context of linear regression indicate that the classical asymptotic HCCME-based tests tend to be liberal.²⁹ However, it remains an open question whether this holds true for the ANCOVA model, too. Moreover, proofs are hardly provided, except in one publication, where the authors nevertheless kept the restrictive assumption of a symmetric error distribution, yet providing some empirical evidence that their approach might work in a more general setting, too.³⁰

1.3 Objectives

The aim of our work is twofold: Firstly, we state a set of assumptions for the general ANCOVA model and prove that they are sufficient for the White HCCME approach²³ to be applied. Henceforth, we will refer to this approach as well as to the associated test statistic by the term White-ANCOVA. Moreover, we discuss what these assumptions actually mean in practice and demonstrate that our approach covers a broad variety of designs which are frequently used in applied research, including multi-way layouts and models with hierarchically nested factors. Hopefully, this detailed discussion will lead to an increased awareness and understanding of the model assumptions, which is of particular importance, since this aspect may have been somewhat neglected in both applied and theoretical work on the general ANCOVA so far. Secondly, we introduce the wild bootstrap method,^{31–33} combine it with the White-ANCOVA and prove the theoretical validity of this approach.

Moreover, we present the results of an extensive simulation study, where we investigate the impact of various degrees of nonnormality and heteroskedasticity on the type I error control of the ANCOVA F test, the HCCME-based test and its wild bootstrap counterpart in several balanced and unbalanced small sample size settings. Such scenarios may well be encountered in practice, but have not been sufficiently examined in the context of ANCOVA so far, not to mention the newly proposed method. Moreover, we simulate the empirical power of the ANCOVA F test and the wild bootstrap version of the HCCME-based test. Finally, we demonstrate the applicability of our method to a real-life dataset and conclude with some discussions of our results, closing remarks and ideas for future research. We deliberately put all mathematical proofs into the Online Appendix (Section 1), in order to ensure that the main body of the paper can be understood by a broad audience. In Sections 2 to 6 of the Online Appendix, we provide additional simulation results for scenarios with alternative HCCMEs as well as for several settings that do not address the main focus of the present paper, but nevertheless provide useful empirical evidence regarding the scope of potential applications (e.g., random covariates, or more severe heteroskedasticity). Moreover, the R code for the simulations and the real-data example is also part of the online supplementary material to this paper.

2 White’s HCCME in the one-way ANCOVA model

In the sequel, we shall denote a diagonal matrix $diag(a_1, a_2, \dots, a_n)$ by the more compact notation $\bigoplus_{i=1}^n a_i$, with an analogous notation for a block diagonal matrix (i.e., a matrix consisting of matrices as diagonal elements). At first, we introduce the HCCME concept. Let us consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{Y} = (Y_1, \dots, Y_N)'$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_c)'$. Moreover, let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$ with $\mathbf{x}_i = (x_i^{(1)}, \dots, x_i^{(c)})'$, $1 \leq i \leq N$, denote the regressor matrix, considered as fixed in the sequel. The errors are assumed to be independent, with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma_i^2 > 0$, $1 \leq i \leq N$. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ denote the ordinary least squares estimator of $\boldsymbol{\beta}$. Moreover, we define

$$\widehat{Cov}(\sqrt{N}\hat{\boldsymbol{\beta}}) := (\mathbf{X}'\mathbf{X}/N)^{-1}N^{-1}\mathbf{X}'\left(\bigoplus_{i=1}^N u_i^2\right)\mathbf{X}(\mathbf{X}'\mathbf{X}/N)^{-1} \tag{2}$$

where $u_i^2 := (Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2$, $1 \leq i \leq N$.

In order to test $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{H} \in \mathbb{R}^{q,c}$, $rank(\mathbf{H}) = q$, White considered a Wald-type test statistic and proved that under certain assumptions, which will be precisely stated in Online Appendix 1, the following asymptotic result holds²³

$$N\{\mathbf{H}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}'\left\{\mathbf{H}\widehat{Cov}(\sqrt{N}\hat{\boldsymbol{\beta}})\mathbf{H}'\right\}^{-1}\mathbf{H}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \chi_q^2 \tag{3}$$

Hence, an asymptotic level α test can be obtained by rejecting the null hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$ if and only if the value of the test statistic

$$T(\mathbf{H}) := N\{\mathbf{H}\hat{\boldsymbol{\beta}}\}'\left\{\mathbf{H}\widehat{Cov}(\sqrt{N}\hat{\boldsymbol{\beta}})\mathbf{H}'\right\}^{-1}\mathbf{H}\hat{\boldsymbol{\beta}} \tag{4}$$

exceeds the $1 - \alpha$ quantile of the central Chi-square distribution with $q = rank(\mathbf{H})$ degrees of freedom. In previous papers, it has been recognized that using White’s initial estimator, as defined in (2), makes the corresponding test statistics liberal.³⁴ Consequently, some refinements of White’s initial estimator had been proposed. For example, one could replace the squared residuals u_i^2 in (2) by $u_i^2/(1 - p_{ii})^{34}$ or $u_i^2/(1 - p_{ii})^{\delta_i}$, $\delta_i := \min(4, p_{ii}/(N^{-1} \sum_{j=1}^N p_{ij}))$,³⁵ where p_{ii} denotes the i -th diagonal element of the hat matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $1 \leq i \leq N$. Note that regarding the proofs provided in Online Appendix 1, this modification does not matter, because $\lim_{N \rightarrow \infty} (1 - p_{ii}) = 1$ and $\lim_{N \rightarrow \infty} (1 - p_{ii})^{\delta_i} = 1$, uniformly in i , under the assumptions stated in Online Appendix 1. Throughout this paper, we refer to White’s initial estimator and its modified versions by HC0, HC2 and HC4, respectively.³⁴

Next, to turn to the general one-way ANCOVA model, suppose that we have $N = n_1 + n_2 + \dots + n_a$ observations of individuals from a different groups. We assume that these observations are realizations of random variables, say, $Y_{11}, Y_{12}, \dots, Y_{an_a}$, which follow the linear model $Y_{ij} = \mu_i + \sum_{k=1}^r v_k z_{ij}^{(k)} + \epsilon_{ij}$, where ϵ_{ij} are independent with $E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma_{ij}^2 > 0$, and $z_{ij}^{(k)}$ are some fixed covariates, $i = 1, \dots, a$,

$j = 1, \dots, n_i, k = 1, \dots, r$. Equivalently, in matrix notation

$$\mathbf{Y} = \left(\bigoplus_{i=1}^a \mathbf{1}_{n_i} \right) \boldsymbol{\mu} + \mathbf{Z}\mathbf{v} + \boldsymbol{\epsilon} \tag{5}$$

where $\mathbf{1}_{n_i}$ denotes the n_i -dimensional vector of ones, $\mathbf{Y} = (Y_{11}, \dots, Y_{an_a})'$, $\mathbf{Z} = (\mathbf{z}_{11}, \dots, \mathbf{z}_{an_a})'$, $\mathbf{z}_{ij} = (z_{ij}^{(1)}, \dots, z_{ij}^{(r)})'$, $1 \leq i \leq a, 1 \leq j \leq n_i$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_a)'$, $\mathbf{v} = (v_1, \dots, v_r)'$, $\boldsymbol{\epsilon} = (\epsilon_{11}, \dots, \epsilon_{an_a})'$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $Cov(\boldsymbol{\epsilon}) = diag(\sigma_{11}^2, \dots, \sigma_{an_a}^2)$. Now, in order to derive a test for

$$H_0 : \mu_1 = \dots = \mu_a \tag{6}$$

we rewrite the ANCOVA model given in (5) in the form of the linear model (1) by setting $\mathbf{X} = (\bigoplus_{i=1}^a \mathbf{1}_{n_i}, \mathbf{Z})$, $\boldsymbol{\beta} = (\boldsymbol{\mu}', \mathbf{v}')'$ and splitting up the indices in (1), in order to account for the grouped structure of the data. Accordingly, in order to express the hypothesis stated in (6) as $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$, we specify $\mathbf{H} = (\mathbf{1}_{a-1}, -\mathbf{I}_{a-1}, \mathbf{0})$ (i.e., \mathbf{H} is a $(a - 1) \times 3$ block matrix, containing the $(a - 1)$ -dimensional vector $\mathbf{1}_{a-1}$ of ones, the $(a - 1)$ -dimensional identity matrix \mathbf{I}_{a-1} multiplied by (-1) and the $((a - 1) \times r)$ matrix containing only zeroes, respectively, where r denotes the number of covariates). Observe that \mathbf{H} has full row rank because $rank(\mathbf{H}) = a - 1$. Furthermore, (6) is indeed equivalent to $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$, because the latter simplifies to $H_0 : \mu_1 - \mu_2 = 0, \dots, \mu_1 - \mu_a = 0$. Now, in the following theorem, we set up some assumptions, which are required for the White HCCME approach to work.

Theorem 1. *In what follows, let \mathbf{I}_a , \mathbf{J}_a and \mathbf{P}_a denote the a -dimensional identity matrix, the a -dimensional square matrix of 1's and the so-called a -dimensional centering matrix (i.e., $\mathbf{P}_a = \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a$), respectively. Let us assume that the following conditions are fulfilled for model (5):*

- (GA1) *The components of the error vector $\boldsymbol{\epsilon}$ are independent, with $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $Cov(\boldsymbol{\epsilon}) = diag(\sigma_{11}^2, \sigma_{12}^2, \dots, \sigma_{1n_1}^2, \dots, \sigma_{a1}^2, \dots, \sigma_{an_a}^2) > 0$, and $\exists d_1 > 0, \gamma > 0 : E(|\epsilon_{ij}|^{2+\gamma}) \leq d_1$ for all $i \in \{1, 2, \dots, a\}, j \in \{1, 2, \dots, n_i\}$.*
- (GA2) *$\exists d_2 > 0 : |z_{ij}^{(k)}| < d_2$ for all $i \in \{1, \dots, a\}, j \in \{1, \dots, n_i\}, k \in \{1, \dots, r\}$.*
- (GA3) (a) *$\exists d_3 > 0, m_0 \in \mathbb{N} : \frac{N}{n_i} \leq d_3$ for all $i \in \{1, 2, \dots, a\}$ and $N \geq m_0$.*
 (b) *$\exists d_4 > 0, m_1 \in \mathbb{N} : det(N^{-1} \sum_{i=1}^a \mathbf{Z}_i' \mathbf{P}_{n_i} \mathbf{Z}_i) = \prod_{s=1}^r \lambda_s \geq d_4$ for all $N \geq m_1$, where \mathbf{Z}_i denotes the regressor matrix of the i -th group, $1 \leq i \leq a$, and $\lambda_1, \dots, \lambda_r$ are the eigenvalues of the matrix $N^{-1} \sum_{i=1}^a \mathbf{Z}_i' \mathbf{P}_{n_i} \mathbf{Z}_i$.*
- (GA4) (a) *$\exists d_5 > 0, m_2 \in \mathbb{N} : d_5 \sum_{j=1}^{n_i} \sigma_{ij}^2 \geq N$ for all $i \in \{1, 2, \dots, a\}$ and $N \geq m_2$.*
 (b) *Let $\mathbf{M} := N^{-1} \sum_{i=1}^a \mathbf{Z}_i' \{ \sum_i - (\sum_{j=1}^{n_i} \sigma_{ij}^2)^{-1} \mathbf{s}_i \mathbf{s}_i' \} \mathbf{Z}_i$, where $\sum_i = diag(\sigma_{i1}^2, \dots, \sigma_{in_i}^2)$, $\mathbf{s}_i = (\sigma_{i1}^2, \dots, \sigma_{in_i}^2)'$, $1 \leq i \leq a$. Then, $\exists d_6 > 0, m_3 \in \mathbb{N} : det(\mathbf{M}) = \prod_{s=1}^r \tau_s \geq d_6$ for all $N \geq m_3$, where $\tau_1, \tau_2, \dots, \tau_r$ denote the eigenvalues of the matrix \mathbf{M} .*

Then, the convergence result (3) holds.

The proof of this theorem is provided in Online Appendix 1. We would like to emphasize that the assumptions (GA1)-(GA4) are very general, because they either exclude trivial cases or impose only weak assumptions on the covariates and the errors, which are met in virtually any real-life setting. In particular, observe that the error distributions may even vary between subjects. So, all in all, the proposed method is potentially useful for a broad range of applications. This will be further illustrated in the following section.

3 Applicability of the White-ANCOVA model to real-life data

At first, we would like to explain how the assumptions (GA1)-(GA4) can be interpreted. Therefore, we shall discuss the assumptions stated in Theorem 1 point by point now.

- (GA1) The errors ϵ_{ij} are required to be independent with expectation 0. These assumptions can be checked by standard techniques (e.g., residual plots). Independence of observations is most likely justifiable, unless, for example, some patients undergo a certain examination in the same room/at the same time point.

The moment condition as stated in (GA1) holds, for example, for errors from a symmetric distribution, but also for any distribution with finite skewness, which is uniformly bounded. Thus, our model is not very restrictive, allowing for a broad range of distributions, including χ^2 , $t(df > 3)$, exponential, Γ and other families of distributions. Even more generally, the case of subject-specific error distributions is covered, as long as their third moments are uniformly bounded. A typical example would be that the distributions of the errors are allowed to vary between groups, not only in terms of variances, but also in terms of the distribution families they belong to.

(GA2) The uniform boundedness condition on the covariates looks rather restrictive at first sight. However, in real-life settings, most variables do actually satisfy this assumption, because their range of values is indeed bounded for obvious reasons (e.g., blood pressure measurements are neither negative nor arbitrarily large).

(GA3) Assumption (a) states that the group sizes n_i , $1 \leq i \leq a$, essentially grow at the same rate (i.e., are not extremely unbalanced). This assumption is usually needed when doing inferential asymptotics. Apart from this formal aspect, the condition also makes sense from a practical point of view: If this assumption was not met, we would, at some point, arrive at a severe underrepresentation of a certain group, and such a case should be excluded.

To turn to (b), if the covariates had been random, $N^{-1} \sum_{i=1}^a \mathbf{Z}_i' \mathbf{P}_{n_i} \mathbf{Z}_i$ would have been the pooled empirical covariance matrix of the covariates. Now, assumption (b) states that this matrix must not be singular for sufficiently large N . If the empirical covariance matrix was not regular for sufficiently large N , this would mean that the variance of either at least one of the covariates or a linear combination of several covariates would converge to 0. Thus, that covariate (or the linear combination) would have, asymptotically, a point-mass distribution, which does not make sense from a practical point of view. So, even in the more general random covariate scenario, it is reasonable to assume (b). When dealing with fixed covariates, a singular 'empirical covariance matrix' would mean that either the values of one covariate did not differ between subjects or some covariates were linearly dependent, which would clearly render inclusion of that covariate (these covariates) into the analysis useless.

(GA4) Assumption (a) is met if, for example, the a averages of the error variances within the groups, $n_i^{-1} \sum_{j=1}^{n_i} \sigma_{ij}^2$, do not become too small if N goes to infinity. To see this, observe that under (GA3)(a), we get $N^{-1} \sum_{j=1}^{n_i} \sigma_{ij}^2 = n_i / N (n_i^{-1} \sum_{j=1}^{n_i} \sigma_{ij}^2) \geq d_3^{-1} (n_i^{-1} \sum_{j=1}^{n_i} \sigma_{ij}^2)$, $1 \leq i \leq a$. So, if the latter sum can be uniformly bounded from below for sufficiently large sample sizes, this would immediately yield that assumption (a) is met. Obviously, such an assumption is sensible because if at least in one particular group, the average of the error variances in that group would go to 0, the regression part of the ANCOVA would not make sense due to the quasi deterministic linear relationship in that group.

To turn to (b), for sake of simplicity, we assume homogeneity of the variances within the groups, that is, $\sigma_{ij}^2 = \sigma_i^2$ for all $j \in \{1, 2, \dots, n_i\}$, $1 \leq i \leq a$. Then, the matrix from (GA4)(b) can be rewritten as

$$M = \frac{1}{N} \sum_{i=1}^a \mathbf{Z}_i' \left\{ \Sigma_i - \left(\sum_{j=1}^{n_i} \sigma_{ij}^2 \right)^{-1} \mathbf{s}_i \mathbf{s}_i' \right\} \mathbf{Z}_i = \frac{1}{N} \sum_{i=1}^a \sigma_i^2 \mathbf{Z}_i' \mathbf{P}_{n_i} \mathbf{Z}_i$$

Therefore, as long as neither all group-specific error variances nor all within-group variances of at least one covariate or a linear combination of several covariates are too close to 0, assumption (GA4)(b) is met.

To close this section, we would like to demonstrate that our model covers a broad range of designs frequently encountered in practice. In order to keep the notation compact, we use the unique projection matrix $\mathbf{T} = \mathbf{H}'(\mathbf{H}\mathbf{H}')^{-1}\mathbf{H}$ to formulate hypotheses about the vector $\boldsymbol{\mu}$ of adjusted means. Note that $\mathbf{T}\boldsymbol{\beta} = \mathbf{0} \Leftrightarrow \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$, so, basically, the only change which has to be made is to replace \mathbf{H} by \mathbf{T} in (3) and take the Moore-Penrose inverse instead of the classical inverse of the covariance matrix. Observe that the asymptotic result (3) still holds, because the corresponding theorems concerning the distribution of quadratic forms are also valid for the Moore-Penrose inverse.³⁶ Furthermore, due to the fact that $\mathbf{H} = (\mathbf{H}_f, \mathbf{0})$, where \mathbf{H}_f denotes the hypothesis matrix corresponding to the factorial part of the parameter vector $\boldsymbol{\beta} = (\boldsymbol{\mu}', \mathbf{v}')'$, the corresponding projection matrix is a block diagonal matrix

of the form $\mathbf{T} = \text{diag}(\mathbf{T}_f, \mathbf{0})$. Now, we briefly sketch how hypotheses about factor effects can be tested in several practically important designs. In what follows, let \mathbf{I}_a , \mathbf{J}_a and \mathbf{P}_a denote the a -dimensional identity matrix, the a -dimensional square matrix of 1's and the so-called a -dimensional centering matrix (i.e., $\mathbf{P}_a = \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a$), respectively.

- **One-way layout.** Suppose you have observations of subjects in a groups (e.g., treatment arms in a clinical trial). The hypothesis (6), that is, the null hypothesis of no difference between the adjusted means, can be formulated by setting $\mathbf{T}_f = \mathbf{P}_a$.
- **Crossed two-way layout with interactions.** Suppose there are two cross-classified fixed factors B and C with levels $i = 1, \dots, b$ and $j = 1, \dots, c$ (e.g., the levels of B could represent different drugs, whereas the levels of C indicate several dosages, which are required to be the same for all drugs). So, the total number of factor level combinations is $a = bc$, and by splitting up the indices, we have $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \dots, \mu_{bc})'$. Using an additive notation, $\mu_{ij} = \mu + v_i + \tau_j + (\nu\tau)_{ij}$, with the usual side conditions $\sum_i v_i = \sum_j \tau_j = \sum_i (\nu\tau)_{ij} = \sum_j (\nu\tau)_{ij} = 0$. The hypotheses of no main effects of the factors B (i.e., $v_i = 0\forall i$) and C (i.e., $\tau_j = 0\forall j$) can be specified by $\mathbf{T}_f = \mathbf{P}_b \otimes \frac{1}{c}\mathbf{J}_c$ and $\mathbf{T}_f = \frac{1}{b}\mathbf{J}_b \otimes \mathbf{P}_c$, respectively. The hypothesis of no interaction effect (i.e., $(\nu\tau)_{ij} = 0\forall i, j$) is given by $\mathbf{T}_f = \mathbf{P}_b \otimes \mathbf{P}_c$.
- **Hierarchically nested two-way design.** By contrast to the design above, assume now that C is nested under B (e.g., for each of the drugs $i = 1, \dots, b$, there are specific dosages $j = 1, \dots, c_i$ being administered). In this design, the vector of adjusted means is $\boldsymbol{\mu} = (\mu_{11}, \dots, \mu_{1c_1}, \dots, \mu_{b1}, \dots, \mu_{bc_b})'$. In additive notation, we can write $\mu_{ij} = \mu + v_i + \tau(v)_{j(i)}$, with the side conditions $\sum_i v_i = \sum_j \tau(v)_{j(i)} = 0$. Then, the hypothesis of no category effect B (i.e., $v_i = 0\forall i$) and sub-category effect $C(B)$ (i.e., $\tau(v)_{j(i)} = 0\forall i, j$) can be formulated via $\mathbf{T}_f = \mathbf{P}_b \otimes \tilde{\mathbf{J}}_c$ and $\mathbf{T}_f = \tilde{\mathbf{P}}_c$, respectively. Thereby, $\tilde{\mathbf{J}}_c = \text{diag}(\frac{1}{c_1}\mathbf{J}_{c_1}, \dots, \frac{1}{c_b}\mathbf{J}_{c_b})$ and $\tilde{\mathbf{P}}_c = \text{diag}(\mathbf{P}_{c_1}, \dots, \mathbf{P}_{c_b})$.

The generalization to factorial designs with more than two cross-classified or nested factors works analogously and is, therefore, not discussed here.

4 The Wild Bootstrap for the White-ANCOVA model

Especially in small sample size scenarios, the White-ANCOVA test statistic and the corresponding asymptotic result stated in (3) might not yield satisfactory results in terms of maintaining the prespecified type I error probability, see our simulation study in Section 5 below. A resampling procedure such as the bootstrap might remedy this problem. In the context of heteroskedastic regression, various so-called wild bootstrap methods have been proposed.^{31–33} The key idea of the wild bootstrap is as follows: Let u_i^2 denote the i -th squared residual of the linear model (1), $1 \leq i \leq N$. Furthermore, let p_{ii} denote the i -th diagonal element of the hat matrix $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $1 \leq i \leq N$. Now, we repeatedly draw random samples consisting of N observations $Y_i^* = \epsilon_i^*$, where $\epsilon_i^* = u_i(1 - p_{ii})^{-1/2}T_i$, $1 \leq i \leq N$, $(T_i)_{i=1}^N$ are i.i.d. and independently generated from the original data, with $E(T_1) = 0$ and $Var(T_1) = 1$. Although for generating the T_i 's, one may choose any distribution which satisfies the latter two conditions, some particular choices have become popular. In this paper, we use the Rademacher distribution, which is defined by $P(T_1 = -1) = P(T_1 = 1) = 1/2$.

For each vector $\mathbf{Y}^* = (Y_1^*, \dots, Y_N^*)'$ of bootstrap observations, we calculate the bootstrap OLS estimate $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^*$ and the bootstrap version of White's covariance matrix estimator (2), that is

$$\widehat{\boldsymbol{\Sigma}}^* := (\mathbf{X}'\mathbf{X}/N)^{-1}N^{-1}\mathbf{X}'\left(\bigoplus_{i=1}^N u_i^{*2}\right)\mathbf{X}(\mathbf{X}'\mathbf{X}/N)^{-1} \tag{7}$$

where $u_i^{*2} := (Y_i^* - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^*)^2$, $1 \leq i \leq N$. Finally, we calculate the bootstrap analogon of White's Wald-type test statistic (4), namely

$$T^*(\mathbf{H}) := N\{\mathbf{H}\hat{\boldsymbol{\beta}}^*\}'\{\mathbf{H}\widehat{\boldsymbol{\Sigma}}^*\mathbf{H}\}^{-1}\mathbf{H}\hat{\boldsymbol{\beta}}^* \tag{8}$$

To turn to the White-ANCOVA setting, we rewrite the ANCOVA model (5) as a special case of the linear model (1), as we have already outlined in Section 2. The main idea of any bootstrap procedure is to resemble the process underlying the generation of the original data reasonably well. In the following theorem, we state that

given the data, the distribution of the bootstrap test statistic (8) indeed mimics the distribution of the original test statistic (4) under the null hypothesis.

Theorem 2. *Let us assume that model (5) as well as the assumptions (GA1)–(GA4) stated in Theorem 1 hold. Let $P_{H_0}(T(\mathbf{H}) \leq x)$ denote the unconditional CDF of $T(\mathbf{H})$ under H_0 and $P_{\beta}(T^*(\mathbf{H}) \leq x|\mathbf{Y})$ the conditional CDF of $T^*(\mathbf{H})$ if $\beta \in \mathbb{R}^{a+r}$ is the true underlying parameter. Then, the following statements hold for any $\beta \in \mathbb{R}^{a+r}$.*

- (1) $\sup_{x \in \mathbb{R}} \left| P_{\beta}(T^*(\mathbf{H}) \leq x|\mathbf{Y}) - \chi_q^2(-\infty, x] \right| \xrightarrow{P} 0$ in probability, where $q = r(\mathbf{H})$.
- (2) $\sup_{x \in \mathbb{R}} \left| P_{\beta}(T^*(\mathbf{H}) \leq x|\mathbf{Y}) - P_{H_0}(T(\mathbf{H}) \leq x) \right| \xrightarrow{P} 0$ in probability.

The proof of this theorem is given in Online Appendix 1. Note that there, we show that in fact, the wild bootstrap test statistic (8) yields an asymptotically valid test in any heteroskedastic linear model under very weak assumptions, which are stated in Online Appendix 1.

5 Simulation study

In order to evaluate the finite-sample performance of our proposed tests, we conducted an extensive simulation study, using R version 3.3.1.³⁷ We assessed the maintenance of a prespecified alpha level of 5%. Hereby, we considered an ANCOVA model with four groups and small to moderate sample sizes, namely $(n_1, n_2, n_3, n_4) \in \{(40, 40, 40, 40), (15, 15, 15, 15), (5, 5, 5, 5), (5, 10, 20, 25), (25, 20, 10, 5)\}$. We assumed two fixed covariate vectors $\mathbf{z}_1, \mathbf{z}_2$. The first one consisted of equally spaced values between -10 and 10 . For the second vector, the first and the second half of the components were equally spaced in $[0, 5]$ and $[-2, -1]$, respectively, sorted in descending order. The regression coefficients corresponding to the two covariates were assumed to be -0.5 and 1.5 , respectively. The vector μ of the group means was set to $\mathbf{0}$, in order to represent an instance of the null hypothesis $H_0 : \mu_1 = \dots = \mu_4$.

For each of the sample size scenarios from above, the errors were drawn from the standard normal, χ_5^2 , lognormal, or double exponential distribution. If required, these errors were appropriately shifted and/or scaled and subsequently multiplied with the square root of the covariance matrix $\bigoplus_{i=1}^a \sigma_i^2 \mathbf{I}_{n_i}$, in order to make sure that the variances of the error terms were indeed equal to the values specified as follows: For the group-wise error variances, we considered the homoskedastic case $\sigma_i^2 = 1$ (Scenario I) as well as the heteroskedastic setting $\sigma_i^2 = i$, $i \in \{1, 2, 3, 4\}$ (Scenario II). Note that although we derived the White-ANCOVA tests under the more general assumption of subject-specific error variances, such a case would hardly be encountered in practice. Most reasonable studies are designed such that the residual variances are rather homogeneous within groups. If this is not the case, it is difficult to interpret the results of the ANCOVA meaningfully. Nevertheless, in order to examine the performance of the White-ANCOVA tests in a more general setting, we also simulated a scenario where within the first group, we assumed a variance of one for the subjects $j = 1, 2, \dots, \lfloor n_1/2 \rfloor$ and a variance of two for the remaining ones, respectively. For the other three groups, we set $\sigma_i^2 = i + 1$, where $i = 2, 3, 4$. This allocation scenario will be referred to as Scenario III.

Finally, the simulated observations were generated according to (5). For each of the 60 scenario combinations, we repeated the data generation process 10,000 times, which yields a standard error of 0.22% for the type I error level $\alpha = 5\%$. Within each simulation run, we drew 5000 bootstrap samples, according to the procedure described in Section 4.

In addition to the White-ANCOVA test statistic and its wild bootstrap version, we also considered the classical ANCOVA F test as a third competitor. It should be noted that the HC4 covariance matrix estimator (see Section 2) was used for both the White-ANCOVA test statistic and its wild bootstrap version. However, we also repeated the simulations for all scenarios using HC0 and HC2, respectively. The results of the latter can be found in Online Appendix 2. Table 1 contains the simulation results for the HC4 estimator.

The White-ANCOVA test tended to be less liberal when it was based on the HC4 estimator instead of the HC0 or the HC2 estimator, whereas the performances of the respective bootstrap versions were similar to each other. Therefore, the following discussion is focused only on the HC4-based tests. In balanced group size scenarios, the classical ANCOVA maintained the prespecified 5% level. In the unbalanced settings, the classical ANCOVA was hardly affected by nonnormality. However, heteroskedasticity led to either substantially deflated or inflated type I error rates, depending on the relation between the variances and the group sizes. In case of positive pairing (i.e., the smaller groups have the smaller variances), the ANCOVA F test tended to be conservative, whereas negative

Table 1. Empirical type I error rates (in %) for the ANCOVA F test, the White-ANCOVA test, and its wild bootstrap version (based on the HC4 estimator).

| Var | N | Standard normal | | | Standard lognormal | | | Double exponential | | | Chi-squared (df = 5) | | |
|-----|-------|-----------------|-------|-----|--------------------|-------|-----|--------------------|-------|-----|----------------------|-------|-----|
| | | F test | White | WB | F test | White | WB | F test | White | WB | F test | White | WB |
| I | n_1 | 5.0 | 6.2 | 5.0 | 4.5 | 4.9 | 5.1 | 5.1 | 6.2 | 5.1 | 4.8 | 5.9 | 5.0 |
| | n_2 | 5.3 | 9.2 | 5.5 | 4.9 | 5.6 | 3.7 | 5.1 | 8.4 | 5.2 | 5.0 | 8.0 | 4.9 |
| | n_3 | 4.5 | 16.6 | 5.9 | 4.2 | 9.9 | 3.1 | 4.9 | 15.9 | 6.4 | 5.0 | 15.7 | 5.7 |
| | n_4 | 5.3 | 8.2 | 4.8 | 4.9 | 5.9 | 3.8 | 4.7 | 7.5 | 4.7 | 5.1 | 8.3 | 5.1 |
| | n_5 | 4.8 | 8.2 | 5.0 | 4.8 | 4.9 | 3.1 | 4.9 | 7.7 | 4.9 | 5.2 | 8.1 | 4.5 |
| II | n_1 | 4.5 | 6.3 | 5.1 | 4.5 | 5.4 | 5.7 | 4.9 | 6.4 | 5.3 | 4.4 | 6.2 | 5.0 |
| | n_2 | 4.8 | 8.9 | 5.3 | 4.7 | 6.2 | 4.3 | 4.9 | 8.4 | 5.4 | 4.9 | 8.4 | 5.0 |
| | n_3 | 4.5 | 16.5 | 5.8 | 4.1 | 10.1 | 3.4 | 5.0 | 16.1 | 6.3 | 4.9 | 15.7 | 5.5 |
| | n_4 | 3.2 | 8.1 | 4.8 | 3.2 | 6.3 | 4.7 | 3.0 | 7.5 | 4.8 | 3.3 | 8.1 | 5.2 |
| | n_5 | 10.1 | 8.1 | 5.0 | 8.2 | 4.8 | 3.2 | 10.0 | 7.9 | 4.9 | 10.5 | 8.2 | 4.6 |
| III | n_1 | 4.8 | 6.3 | 5.1 | 4.8 | 5.4 | 5.6 | 5.0 | 6.4 | 5.4 | 4.6 | 6.1 | 5.0 |
| | n_2 | 5.1 | 9.1 | 5.4 | 4.9 | 6.0 | 4.3 | 5.0 | 8.6 | 5.3 | 4.9 | 8.4 | 5.0 |
| | n_3 | 4.7 | 16.8 | 5.7 | 4.2 | 10.2 | 3.1 | 5.1 | 16.2 | 6.3 | 5.0 | 16.0 | 5.5 |
| | n_4 | 3.5 | 8.1 | 4.8 | 3.4 | 6.1 | 4.2 | 3.3 | 7.4 | 4.7 | 3.5 | 8.0 | 5.0 |
| | n_5 | 9.5 | 8.2 | 5.1 | 7.9 | 4.9 | 3.1 | 9.5 | 7.9 | 4.9 | 10.0 | 8.3 | 4.5 |

I: $\sigma_1^2 = \dots = \sigma_4^2 = 1$, II: $\sigma_i^2 = i, i \in \{1, 2, 3, 4\}$, III: $\sigma_{1j}^2 = \dots = \sigma_{ij}^2 = 1$ and $\sigma_{(j+1)}^2 = \dots = \sigma_{in_j}^2 = 2, j = \lfloor n_1/2 \rfloor, \sigma_i^2 = i + 1$ for group $i, i \in \{2, 3, 4\}$. $n_1 = (40, 40, 40, 40), n_2 = (15, 15, 15, 15), n_3 = (5, 5, 5, 5), n_4 = (5, 10, 20, 25), n_5 = (25, 20, 10, 5)$. The data generation process was repeated 10,000 times. For each simulation run, 5000 bootstrap samples were generated.

pairing (i.e., the smaller groups have the larger variances) made the test liberal, as suggested by conventional wisdom. These effects became more pronounced when the differences between the group variances were increased (see Online Appendix 4). The asymptotic White-ANCOVA test yielded substantially inflated type I error rates, both in homo- and heteroskedastic settings. Clearly, the wild bootstrap version maintained the prespecified level well in all scenarios, being superior to the ANCOVA F test in terms of type I error control especially in case of heteroskedasticity and unequal group sizes. The slight conservatism seen for lognormal errors might be caused by the underlying method of estimating the covariance matrix, since for the White-ANCOVA, we also observed lower type I error rates in the lognormal case, compared to the other distributions. Indeed, a closer empirical examination revealed that the variances of the White covariance matrix estimator and its wild bootstrap counterpart were larger for lognormal errors than for the other distributions under consideration (for details, see Online Appendix 6).

Subsequently, we compared the aforementioned tests with respect to their empirical power. However, we only considered the ANCOVA F test and the wild bootstrap version of the White-ANCOVA, because the asymptotic White-ANCOVA showed a poor performance in terms of maintaining the type I error rates. Furthermore, in order to make sure that the prespecified level was maintained by both tests, we only considered a homoskedastic, balanced setting with $\sigma_1^2 = \sigma_2^2 = 1$ and $n_1 = n_2 = 15$. Moreover, we specified fixed alternatives by setting $\mu_1 = 0, \mu_2 = \delta$, where $\delta \in \{0, 0.1, \dots, 3.0\}$. The four error distributions were chosen as described above. For each scenario, we conducted 10,000 simulations and 5000 bootstrap runs, respectively. The results are displayed in Figure 1.

For small values of δ , the wild bootstrap test appeared to be more powerful than the classical ANCOVA. As δ increased, this relationship was gradually being reversed. However, the power of the ANCOVA F test at most exceeded the empirical power of the wild bootstrap test by six to seven percentage points. So, the bootstrap version of the White-ANCOVA never suffered from a substantial power loss compared to the classical ANCOVA test, even when the assumptions of the latter were met.

Although for ease of presentation, the case of random covariates was not formally considered in the present paper, some simulation results are provided in Online Appendix 3. All specifications were the same as in the fixed covariate setting, with the only difference that the observations of the second covariate were generated from one of the following distributions: the uniform distribution on $[0, 10]$, the standard normal, the standard lognormal, and the *Poisson*($\lambda = 5$) distribution. Overall, the results were very similar to those from the fixed covariate settings. The bootstrap-based method performed even slightly better, being less conservative in case of lognormal errors.

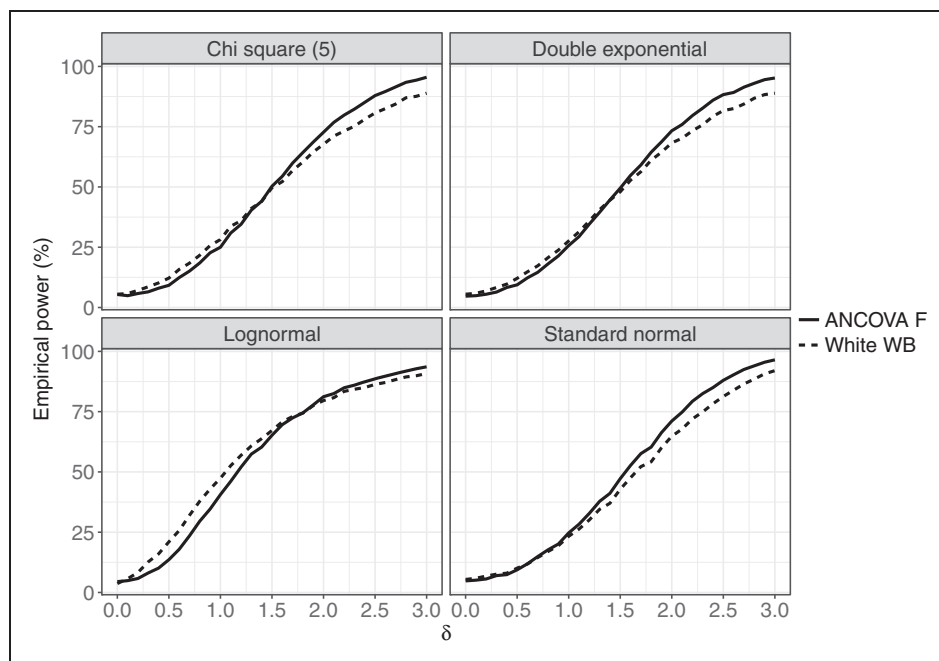


Figure 1. Empirical power for the ANCOVA F test (solid) and the wild bootstrap version of the White-ANCOVA test (HC4 version; dashed). Data were generated for two groups with $\mu_1 = 0$, $\mu_2 = \delta$, $\sigma_1^2 = \sigma_2^2 = 1$, $n_1 = n_2 = 15$.

Moreover, the maximum power loss of the bootstrap procedure compared to the classical ANCOVA was slightly smaller than in the fixed covariate setting. It should be noted that both with fixed and with random covariates, the somewhat low empirical type I error rates in case of lognormal errors did not translate to an overall loss of power. In fact, for popular choices of the target power (e.g., 80% or 90%), the bootstrap-based method had basically the same if not even a slightly better performance (in case of random covariates) compared to the ANCOVA F test (see the lower left panels of Figures 1 to 4 in Section 3 of the Online Appendix).

Finally, we also considered some rather extreme scenarios, in order to examine potential limitations of our proposed method. To begin with, especially for the practitioners, it is of interest to give at least some rough advice about sample sizes which are too small for use of the bootstrap-based approach. It is obvious from the simulation results reported in Section 5 of the Online Appendix that our proposed method tended to yield liberal results for total sample sizes of 15 and below. In homoskedastic settings as well as for balanced heteroskedastic scenarios, the ANCOVA F test is recommended. In cases where the group sizes are 5 and 10 in groups 1 and 2, respectively, and heteroskedasticity might be present, neither of the three tests stayed close to the target 5% level. Secondly, we examined the behaviour of the three methods under consideration when the group allocation ratio was more extreme. Again, from Table 8 in the Online Appendix, we clearly see that the effects of positive and negative pairing on the ANCOVA F test were becoming more and more pronounced. The asymptotic White approach and its bootstrap analogue still performed well, with the latter being slightly closer to the target type I error level. Observe, however, that as the group allocation ratio increased (i.e., for $n_1:n_2 = 1:8$ or even $1:16$), there was some tendency towards either conservative or liberal results, depending on whether positive or negative pairing was present. Still, both White-based approaches outperformed the ANCOVA F test. We would like to mention that from a practical point of view, commonly used allocation ratios are less extreme (e.g., 1:2, 1:3). So, actually, the case of severe group size imbalance might be more interesting from a theoretical rather than from an applied researcher's perspective, unless in cases when, for example, the ANCOVA is used for analysing data from observational studies, where there might be large differences in subgroup sizes.

6 Real-life data example

We illustrate the theoretical considerations from the previous sections by applying the White-ANCOVA as well as its wild bootstrap counterpart to a dataset from a preclinical study of the urological research group and the

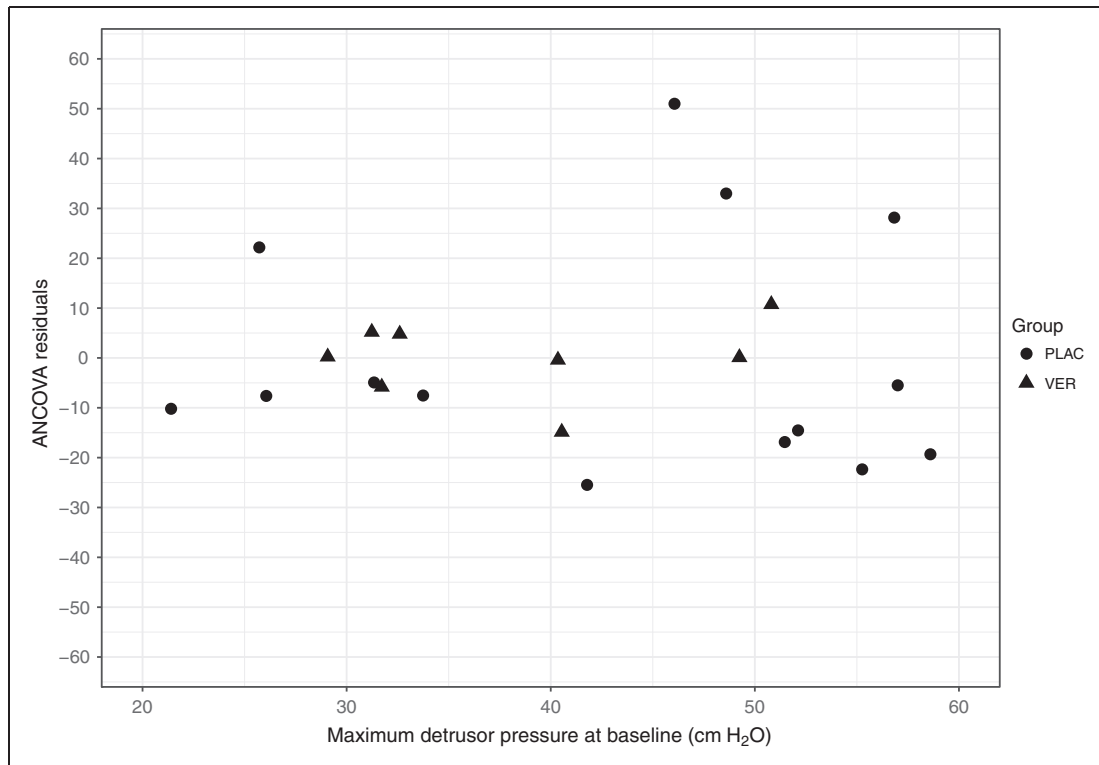


Figure 2. Residual plot for the analysis of covariance, with the change in maximum detrusor pressure from baseline to four weeks post-SCI as outcome and baseline maximum detrusor pressure as a covariate.

Institute of Molecular Regenerative Medicine of the Spinal Cord Injury and Tissue Regeneration Center Salzburg (SCI-TReCS), Paracelsus Medical University, Salzburg, Austria. The aim of this study was to assess the efficacy of an anti-inflammatory drug in a rat model of complete SCI. After SCI, the bladder is known to turn into a pathological state, as indicated by alterations in cystometric variables, such as the voided volume and detrusor pressure.³⁸ We would like to emphasize that these pathological changes in human urinary bladder function are among the most serious consequences of SCI, being associated with a substantially decreased quality of life and a high risk of mortality.^{39,40} Therefore, the main goal of any pharmacological treatment in SCI is to improve cystometric characteristics towards the pre-SCI level. However, the low prevalence of SCI most likely translates to small sample sizes in studies on SCI patients. Moreover, in preclinical research, it is desirable to sacrifice only a small number of animals, due to ethical reasons. Therefore, in research on SCI, or more generally, on any rare disease, statistical methods which perform well in small sample sizes are much needed.

We analyse only a subset of the data that was collected. In the sequel, we shall consider a randomized, parallel two-group design, where the rats received either verum (VER, $N=8$) or placebo (PLAC, $N=14$). The anti-inflammatory drug or the placebo was administered daily, starting at the day of SCI at a standard dosage. Measurements of the urinary bladder function (besides other parameters the maximum detrusor pressure during voiding in cmH_2O and the voided volume in ml) were taken prior to SCI (baseline) and at 1, 7, 14, 21 and 28 days post injury. For ease of illustration, we shall only consider the maximum detrusor pressure (P_{det}) in the sequel. At each time point, P_{det} was obtained as the average over all micturitions within a single time slot of 1 hour. In the same way, the baseline value for each rat had been determined. We consider the difference between the maximum detrusor pressure at 28 days post-SCI and baseline as outcome and P_{det} at baseline as a covariate, respectively. This is in line with the recommendations in a recently published EMA guideline.¹

Before we actually conduct the White-ANCOVA and its wild bootstrap counterpart, we take a look at the assumptions (GA1)-(GA4), which have already been discussed with regard to their interpretation in real-life settings in Section 3. To begin with, it is reasonable to assume that the errors have expectation 0, because the means of the within-group residuals are close to 0 ($\bar{u}_{VER} = 2.22 \times 10^{-16}$, $\bar{u}_{PLAC} = 3.17 \times 10^{-16}$). Moreover,

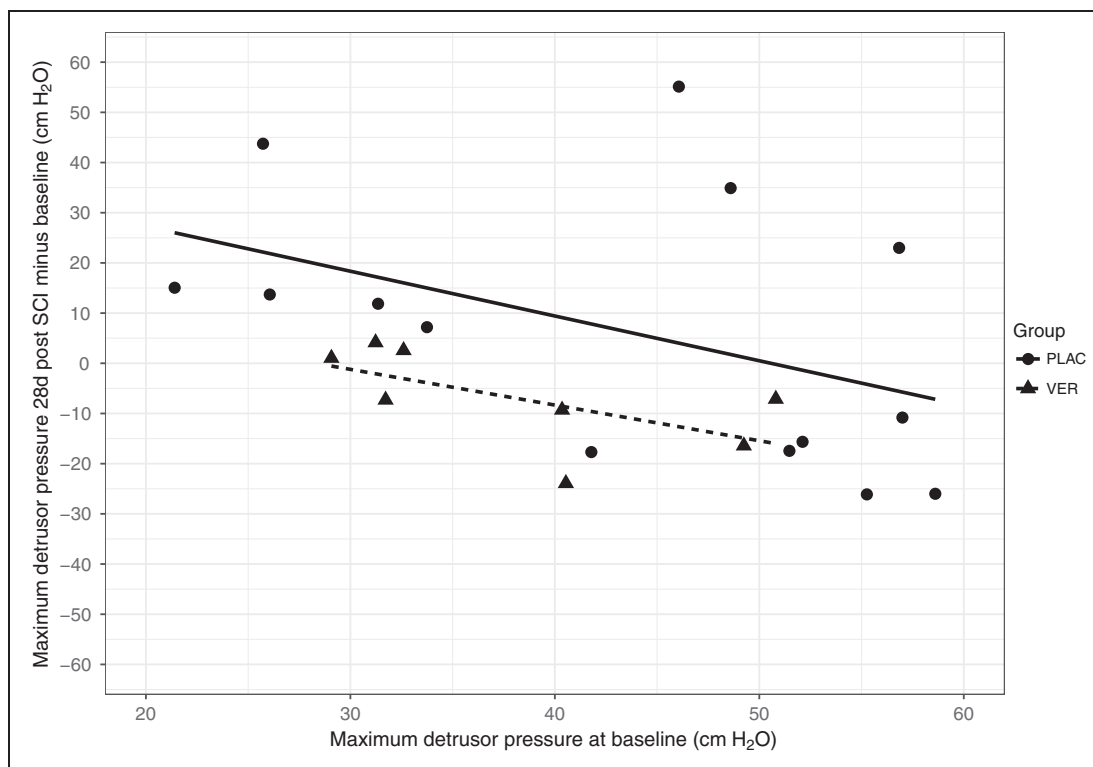


Figure 3. Results from the within-group regression of the change of the maximum detrusor pressure from baseline to four weeks post-SCI on the baseline maximum detrusor pressure.

although the residuals for the PLAC group seem to be skewed (see Figure 2), the moment condition stated in (GA1) is most likely met. Note that the skewness itself is not a problem at all, as long as the skewness is uniformly bounded when the sample size grows. This corresponds to the requirement that the model provides a reasonable fit to the data, which can be ensured by, for example, adding covariates or transformations thereof to the model. The independence of the errors, however, cannot be assumed for all subjects, because only one rat each was taken out of ten cages, but two rats each were taken out of six cages. This issue as well as other potential factors related to the laboratory environment have to be carefully considered in virtually any preclinical study. However, as this is not the main focus of the present work, we shall neglect this aspect for now. Nevertheless, it should be emphasized that our proposed method is not capable of accounting for clustered observations. Therefore, it has to be ensured by properly designing the study that clustering is present at most to a negligible extent.

For physiological reasons, we may rightfully assume that (GA2) holds. Assumption (GA3)(a) can be imposed by design, unless unforeseen events in the laboratory environment or in the population occur and lead to a substantial decrease of observations in one of the groups (e.g., sudden death of a large number of rats in one group). Furthermore, from Figures 2 and 3, it is apparent that neither the variance of the baseline maximum detrusor pressure nor the variance of the residuals is 0. So, all in all, the assumptions (GA1)-(GA4) seem justified.

In general, before conducting an ANCOVA, one has to check if the within-group regression slopes are equal. Indeed, for our data, the regression lines are more or less parallel, because $\hat{\beta}_{PLAC} = -0.89$ ($SE = 0.52$) and $\hat{\beta}_{VER} = -0.71$ ($SE = 0.37$) (also see Figure 3). Therefore, we proceed with the data analysis. Clearly, we have a heteroskedastic small sample size setting with positive pairing ($\hat{\sigma}_{VER}^2 = 60.1$, $\hat{\sigma}_{PLAC}^2 = 558.6$). The ANCOVA F test is not appropriate in such a setting, but the assumptions (GA1)-(GA4) are met. Therefore, we shall compare the results of the White-ANCOVA and its wild bootstrap version now.

Firstly, the estimated adjusted means in the placebo and verum group are 43.69 and 25.78, respectively. The White-ANCOVA yields a p value of 0.0078, whereas the p value resulting from the wild bootstrap White-ANCOVA is 0.0133. This is in line with the findings of our simulation study, where the asymptotic White-ANCOVA showed a clear tendency towards liberal test decisions. It should be emphasized again that although both tests are asymptotically valid under one and the same set of assumptions (GA1)-(GA4), they might yield

substantially different results in practice, not only in terms of concrete p values, but also with respect to keeping the prespecified type I error rate in small samples, as demonstrated in Section 5. As this is a serious threat to the validity of the inferential conclusions drawn from the results, caution is apparently needed when applying the White-ANCOVA in heteroskedastic small-sample settings.

7 Concluding remarks

As outlined in Section 1, the classical ANCOVA and its bootstrap counterpart as well as the HCCME-based approach have been used in many applied research disciplines. However, the performance of each of these methods in small samples has not been satisfactory, and their combination has not been systematically studied yet. Moreover, in the context of an ANCOVA model, the assumptions underlying the White HCCME approach have neither been thoroughly examined theoretically nor discussed in detail in applied research so far. In this paper, we have considered a general ANCOVA model and set up the asymptotic White-ANCOVA test statistic as well as its wild bootstrap counterpart. We have proved that under one and the same set of assumptions, both approaches yield asymptotic level α tests. Note that actually, our proof for the wild bootstrap inference does not only cover the ANCOVA, but also the more general case of a heteroskedastic linear model. In contrast to the work of Mammen, who considered an ANOVA-type statistic in a more general case where the model dimension is allowed to vary with the sample size,³³ our proof for the Wald-type statistic uses relatively straightforward techniques. Our proposed method relies on rather weak assumptions which are met in virtually any practical situation. Therefore, it can be utilized in a broad variety of applied research disciplines. Nevertheless, we strongly encourage applied researchers to check the assumptions (GA1)-(GA4) before doing the actual analyses, as illustrated in Section 6.

Moreover, the results of the simulations presented in Section 5 indicate that the direct White-ANCOVA test should not be used in small samples, due to severely inflated type I error rates. However, the wild bootstrap version of the White-ANCOVA showed a similar performance as the classical ANCOVA F test in balanced settings and outperformed the latter when group sizes were not equal. The only slight drawback of our proposed test is that it tends to be a bit conservative for errors from a lognormal distribution. However, all in all, we recommend using the wild bootstrap version of the White-ANCOVA test when group sizes are small and unbalanced. For example, such a situation may well be encountered in studies on rare diseases (e.g., SCI) or in preclinical trials. Moreover, our work might also be of considerable relevance for medical centres of small to moderate size. Conducting a trial with a small sample of subjects could be an appealing alternative as compared to taking part in a multicentre trial, because fewer human and financial resources are needed, although limited generalizability due to smaller sample sizes could remain as an issue.

It should be noted that in the context of heteroskedastic regression models, some authors have considered a ‘restricted bootstrap’ method, where the residuals are calculated for a model under H_0 .²⁹ In the ANCOVA setting, this would mean that at first, a regression model without group indicators is fitted to the data. The estimates and residuals from that model are used for generating the bootstrap observations, then.²⁹ We have not considered this approach in the present paper, since our proposed method might be more straightforward and easier to understand for applied researchers. As already mentioned above, the small-sample performance of the ‘unrestricted’ approach is very good, so we think that our proposed method can be recommended for both statistical and practical reasons.

Future research will be aimed at extending the approach presented here in several directions: On the one hand, we will investigate different heteroskedastic multivariate ANCOVA (MANCOVA) methods, particularly focusing on small-sample performance. Moreover, we want to study ANCOVA for clustered data. In this context, the adaption of the cluster-robust covariance matrix estimation techniques proposed by Cameron et al.⁴¹ together with an examination of their theoretical properties appears to be a promising line of action.

Acknowledgements

The authors thank Esra Keller and Karin Roider (University Clinic of Urology and Andrology, Spinal Cord Injury and Tissue Regeneration Centre, Paracelsus Medical University, Salzburg, Austria) and Ludwig Aigner (Institute of Molecular Regenerative Medicine, Spinal Cord Injury and Tissue Regeneration Centre, Paracelsus Medical University, Salzburg, Austria) for generously granting access to their acquired data, which has been used as an illustrative example in Section 6. Furthermore, the authors are grateful to the Editor, the Associate Editor and two expert reviewers, for their valuable comments, which improved the quality of the present manuscript considerably.


Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclose receipt of the following financial support for the research, authorship, and/or publication of this article: Arne C Bathke was supported by the Austrian Science Fund (FWF; grant I 2697-N31) and Markus Pauly by the German Research Foundation (DFG; project DFG-PA 2409/4-1), both within a joined D-A-CH Lead Agency Project.

ORCID iD

Georg Zimmermann  <http://orcid.org/0000-0002-8282-1034>

Supplemental material

Supplemental material for this article is available online.

References

1. European Medicines Agency. Guideline on adjustment for baseline covariates in clinical trials. EMA/CHMP/295050/2013, London, UK, 2015.
2. Huitema B. *The analysis of covariance and alternatives: statistical methods for experiments, quasi-experiments, and single-case studies*. New York: Wiley, 2011.
3. Bossa M, Zacur E, Olmos S, et al. Statistical analysis of relative pose information of subcortical nuclei: application on ADNI data. *Neuroimage* 2011; **55**: 999–1008.
4. Lu K. An efficient analysis of covariance model for crossover thorough QT studies with period-specific pre-dose baselines. *Pharm Stat* 2014; **13**: 388–396.
5. Keselman H, Huberty C, Lix L, et al. Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Rev Educ Res* 1998; **68**: 350–386.
6. Misra R. A multivariate procedure for comparing mean vectors for populations with unequal regression coefficient and residual covariance matrices. *Biom J* 1996; **38**: 415–424.
7. Adams K, Brown G and Grant I. Analysis of covariance as a remedy for demographic mismatch of research subject groups: some sobering simulations. *J Clin Exp Neuropsychol* 1985; **7**: 445–462.
8. Berman N and Greenhouse S. Adjusting for demographic covariates by the analysis of covariance. *J Clin Exp Neuropsychol* 1992; **14**: 981–982.
9. Owen S and Froman R. Uses and abuses of the analysis of covariance. *Res Nurs Health* 1998; **21**: 557–562.
10. Pocock S, Assmann S, Enos L, et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002; **21**: 2917–2930.
11. Senn S. Change from baseline and analysis of covariance revisited. *Stat Med* 2006; **25**: 4334–4344.
12. Glass G, Peckham P and Sanders J. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev Educ Res* 1972; **42**: 237–288.
13. Ashford J and Brown S. Generalised covariance analysis with unequal error variances. *Biometrics* 1969; **25**: 715–724.
14. Sadooghi-Alvandi S and Jafari A. A parametric bootstrap approach for one-way ANCOVA with unequal variances. *Commun Stat Theory Methods* 2013; **42**: 2473–2498.
15. Koch G, Tangen C, Jung J, et al. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Stat Med* 1998; **17**: 1863–1892.
16. Tangen C and Koch G. Non-parametric analysis of covariance for confirmatory randomized clinical trials to evaluate dose-response relationships. *Stat Med* 2001; **20**: 2585–2607.
17. Lesaffre E and Senn S. A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Stat Med* 2003; **22**: 3583–3596.
18. Bathke A and Brunner E. A nonparametric alternative to analysis of covariance. In: Akritas M and Politis D (eds.) *Recent advantages and trends in nonparametric statistics*. Amsterdam: Elsevier, 2003, pp.109–120.
19. Tsiatis A, Davidian M, Zhang M, et al. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Stat Med* 2008; **27**: 4658–4677.
20. Thas O, Neve J, Clement L, et al. Probabilistic index models. *J R Stat Soc Series B Stat Methodol* 2012; **74**: 623–671.
21. Chausse P, Liu J and Luta G. A simulation-based comparison of covariate adjustment methods for the analysis of randomized controlled trials. *Int J Environ Res Public Health* 2016; **13**: 414.

22. Fan C and Zhang D. Rank repeated measures analysis of covariance. *Commun Stat Theory Methods* 2017; **46**: 1158–1183.
23. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; **48**: 817–838.
24. Zhu T, Liu X, Connelly P, et al. An optimized wild bootstrap method for evaluation of measurement uncertainties of DTI-derived parameters in human brain. *Neuroimage* 2008; **40**: 1144–1156.
25. Barton S, Crozier S, Lillycrop K, et al. Correction of unexpected distributions of p values from analysis of whole genome arrays by rectifying violation of statistical assumptions. *BMC Genomics* 2013; **14**: 161.
26. Kimura D. Testing nonlinear regression parameters under heteroscedastic, normally distributed errors. *Biometrics* 1990; **46**: 697–708.
27. Judkins D and Porter K. Robustness of ordinary least squares in randomized clinical trials. *Stat Med* 2016; **35**: 1763–1773.
28. Hayes A and Cai L. Using heteroskedasticity-consistent standard error estimators in OLS regression: an introduction and software implementation. *Behav Res Methods* 2007; **39**: 709–722.
29. MacKinnon J. Thirty years of heteroskedasticity-robust inference. Queen’s Economic Department Working Paper 1268, Queen’s University, Ontario, Canada, http://qed.econ.queensu.ca/working_papers/papers/qed_wp_1268.pdf (2012, accessed 24 November 2018).
30. Davidson R and Flachaire E. The wild bootstrap, tamed at last. *J Econom* 2008; **146**: 162–169.
31. Wu C. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat* 1986; **14**: 1261–1295.
32. Liu R. Bootstrap procedures under some non-i.i.d. models. *Ann Stat* 1988; **16**: 1696–1708.
33. Mammen E. Bootstrap and wild bootstrap for high dimensional linear models. *Ann Stat* 1993; **21**: 255–285.
34. MacKinnon J and White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J Econom* 1985; **29**: 305–325.
35. Cribari-Neto F. Asymptotic inference under heteroskedasticity of unknown form. *Comput Stat Data Anal* 2004; **45**: 215–233.
36. Ravishanker N and Dey D. *A first course in linear model theory*. New York: Chapman & Hall/CRC, 2002.
37. R Development Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.
38. Mitsui T, Murray M and Nonomura K. Lower urinary tract function in spinal cord-injured rats: midthoracic contusion versus transection. *Spinal Cord* 2014; **52**: 658–61.
39. Craggs M, Balasubramaniam A, Chung E, et al. Aberrant reflexes and function of the pelvic organs following spinal cord injury in man. *Auton Neurosci* 2006; **126–127**: 355–70.
40. Cruz C and Cruz F. Spinal cord injury and bladder dysfunction: new ideas about an old problem. *Scientific World J* 2011; **11**: 214–234.
41. Cameron A, Gelbach J and Miller D. Bootstrap-based improvements for inference with clustered errors. *Rev Econ Stat* 2008; **90**: 414–427.