Communications in Statistics—Simulation and Computation®, 38: 2019–2038, 2009

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910903243687

A Comparison of Analysis of Covariate-Adjusted Residuals and Analysis of Covariance

ELVAN CEYHAN[1] AND CARLA L. GOAD[2]

[1]Department of Mathematics, Koç University, Istanbul, Turkey
[2]Department of Statistics, Oklahoma State University, Stillwater, Oklahoma, USA

Various methods to control the influence of a covariate on a response variable are compared. These methods are ANOVA with or without homogeneity of variances (HOV) of errors and Kruskal–Wallis (K–W) tests on (covariate-adjusted) residuals and analysis of covariance (ANCOVA). Covariate-adjusted residuals are obtained from the overall regression line fit to the entire data set ignoring the treatment levels or factors. It is demonstrated that the methods on covariate-adjusted residuals are only appropriate when the regression lines are parallel and covariate means are equal for all treatments. Empirical size and power performance of the methods are compared by extensive Monte Carlo simulations. We manipulated the conditions such as assumptions of normality and HOV, sample size, and clustering of the covariates. The parametric methods on residuals and ANCOVA exhibited similar size and power when error terms have symmetric distributions with variances having the same functional form for each treatment, and covariates have uniform distributions within the same interval for each treatment. In such cases, parametric tests have higher power compared to the K–W test on residuals. When error terms have asymmetric distributions or have variances that are heterogeneous with different functional forms for each treatment, the tests are liberal with K–W test having higher power than others. The methods on covariate-adjusted residuals are severely affected by the clustering of the covariates relative to the treatment factors when covariate means are very different for treatments. For data clusters, ANCOVA method exhibits the appropriate level. However, such a clustering might suggest dependence between the covariates and the treatment factors, so makes ANCOVA less reliable as well.

Keywords Allometry; ANOVA; Clustering; Homogeneity of variances; Isometry; Kruskal–Wallis test; Linear models; Parallel lines model.

Mathematics Subject Classification 62J10; 62F03; 62J05.

1. Introduction

In an experiment, the response variable may depend on the treatment factors and quite often on some external factor that has a strong influence on the

Received April 11, 2009; Accepted August 4, 2009
Address correspondence to Elvan Ceyhan, Department of Mathematics, Koç University, Sariyer 34450, Istanbul, Turkey; E-mail: elceyhan@ku.edu.tr

response variable. If such external factors are qualitative or discrete, then *blocking* can be performed to remove their influence. However, if the external factors are quantitative and continuous, the effect of the external factor can be accounted for by adopting it as a *covariate* (Kuehl, 2000), which is also called a *concomitant variable* (Milliken and Johnson, 2002; Ott, 1993). Throughout this article, a covariate is defined to be a variable that may affect the relationship between the response variable and factors (or treatments) of interest, but is not of primary interest itself. Maxwell et al. (1984) compared methods of incorporating a covariate into an experimental design and showed that it is not correct to consider the correlation between the dependent variable and covariate in choosing the best technique. Instead, they recommend considering whether scores on the covariate are available for all subjects prior to assigning any subject to treatment conditions and whether the relationship of the dependent variable and covariate is linear.

In various disciplines such as ecology, biology, medicine, etc., the goal is comparison of a response variable among several treatments after the influence of the covariate is removed. Different techniques are used or suggested in statistical and biological literature to remove the influence of the covariate(s) on the response variable (Huitema, 1980). For example, in ecology one might want to compare richness-area relationships among regions, shoot ratios of plants among several treatments, and of C:N ratios among sites (Garcia-Berthou, 2001). There are three main statistical techniques for attaining this goal: (i) analysis of the ratio of response to the covariate; (ii) analysis of the residuals from the regression of the response with the covariate; and (iii) analysis of covariance (ANCOVA).

Analysis of the ratios is perhaps the oldest method used to remove the covariate effect (e.g., size effect in biology) (see Albrecht et al., 1993, for a comprehensive review). Although many authors recommend its disuse (Atchley et al., 1976; Packard and Boardman, 1988), it might still appear in literature on occasion (Albrecht et al., 1993). For instance, in physiological and nutrition research, the data are scaled by taking the ratio of the response variable to the covariate. Using the ratios in removing the influence of the covariate on the response depends on the relationship between the response and the covariate variables (Raubenheimer and Simpson, 1992). Regression analysis of a response variable on the covariate(s) is common to detect such relationships, which are categorized as *isometric* or *allometric* relationships (Small, 1996). *Isometry* occurs when the relationship between a response variable and the covariate is linear with a zero intercept. If the relationship is nonlinear or if there is a non zero intercept, it is called *allometry*. In allometry, the influence of the covariate cannot be removed by taking the ratio of the response to the covariate. In both of allometry and isometry cases, ANOVA on ratios (i.e., response/covariate values) introduces heterogeneity of variances in the error terms. Hence, ANOVA on ratios may give spurious and invalid results for treatment comparisons, so ANCOVA is recommended over the use of ratios (Raubenheimer and Simpson, 1992). See Ceyhan (2000) for a detailed discussion on the use of ratios to remove the covariate influence.

An alternative method to remove the effect of a covariate on the response variable in biological and ecological research is the use of residuals (Garcia-Berthou, 2001). In this method, an overall regression line is fitted to the entire data set and residuals are obtained from this line (Beaupre and Duvall, 1998). Henceforth, these residuals will be referred to as *covariate-adjusted residuals*. This method was recommended in ecological literature by Jakob et al. (1996) who called it "residual

index" method. Then treatments are compared with ANOVA with HOV on these residuals.

Due to the problems associated with the use of ratios in removing the influence of the covariate from the response, ANOVA (with HOV) on covariate-adjusted residuals and ANCOVA were recommended over the use of ratios (Atchley et al., 1976; Packard and Boardman, 1988). For example, Beaupre and Duvall (1998) used ANOVA on covariate-adjusted residuals in a zoological study. Ceyhan (2000) compared the ANCOVA and ANOVA (with HOV) on covariate-adjusted residuals. ANCOVA has been widely applied in ecology and it was shown to be a superior alternative to ratios by Garcia-Berthou (2001) who also pointed out problems with the residual index and recommends ANCOVA as the correct alternative. He also discussed the differences between ANCOVA and ANOVA on the residual index. They argued that the residual analysis is totally misleading as (i) residuals are obtained from an overall regression on the pooled data, (ii) the residual analysis uses the wrong degrees of freedom in inference, and (iii) residuals fail to satisfy the ANOVA assumptions even if the original data did satisfy them. In fact, Maxwell et al. (1985) also demonstrated the inappropriateness of ANOVA on residuals.

Although ANCOVA is a well-established and highly recommended tool, it also has critics. However, the main problem in literature is not the inappropriateness of ANCOVA, rather its misuse and misinterpretation. For example, Rheinheimer and Penfield (2001) investigated how the empirical size and power performances of ANCOVA are affected when the assumptions of normality and HOV, sample size, number of treatment groups, and strength of the covariate-dependent variable relationship are manipulated. They demonstrated that for balanced designs, the ANCOVA F test was robust and was often the most powerful test through all sample-size designs and distributional configurations. Otherwise, it was not the best performer. In fact, the assumptions for ANCOVA are crucial for its use; especially, the independence between the covariate and the treatment factors is an often ignored assumption resulting incorrect inferences (Miller and Chapman, 2001). This violation is very common in fields such as psychology and psychiatry, due to non random group assignment in observational studies, and Miller and Chapman (2001) also suggested some alternatives for such cases. Hence, the recommendations in favor on ANCOVA (including ours) are valid only when the underlying assumptions are met.

In this article, we demonstrate that it is not always wrong to use the residuals. We also discuss the differences between ANCOVA and analysis of residuals, provide when and under what conditions the two procedures are appropriate and comparable. Then under such conditions, we not only consider ANOVA (with HOV), but also ANOVA without HOV and Kruskal–Wallis (K–W) test on the covariate-adjusted residuals. We provide the empirical size performance of each method under the null case and the empirical power under various alternatives using extensive Monte Carlo simulations.

The nonparametric analysis by K–W test on the covariate-adjusted residuals is actually not entirely nonparametric, in the sense that, the residuals are obtained from a fully parametric model. However, when the covariate is not continuous but categorical data with ordinal levels, then a nonparametric version of ANCOVA can be performed (see, e.g., Akritas et al., 2000; Tsangari and Akritas, 2004a). Further, the nonparametric ANCOVA model of Akritas et al. (2000) is extended to longitudinal data for up to three covariates (Tsangari and Akritas, 2004b).

Additionally, there are nonparametric methods such as Quade's procedure, Puri and Sen's solution, Burnett and Barr's rank difference scores, Conover and Iman's rank transformation test, Hettmansperger's procedure, and the Puri–Sen–Harwell–Serlin test which can be used as alternatives to ANCOVA (see Rheinheimer and Penfield, 2001, for the comparison of the these tests with ANCOVA and relevant references). In fact, Rheinheimer and Penfield (2001) showed that with unbalanced designs, with variance heterogeneity, and when the largest treatment-group variance was matched with the largest group sample size, these nonparametric alternatives generally outperformed the ANCOVA test.

The methods to remove covariate influence on the response are presented in Sec. 2, where the ANCOVA method, ANOVA with HOV and without HOV on covariate adjusted residuals, and K–W test on covariate-adjusted residuals are described. A detailed comparison of the methods, in terms of the null hypotheses, and conditions under which the parametric tests are equivalent are provided in Sec. 3. The Monte Carlo simulation analysis used for the comparison of the methods in terms of empirical size and power is provided in Sec. 4. A discussion together with a detailed guideline on the use of the discussed methods is provided in Sec. 5.

2. ANCOVA and Methods on Covariate-Adjusted Residuals

2.1. *ANCOVA Method*

For convenience, only ANCOVA with a one-way treatment structure in a completely randomized design and a single covariate is investigated. A simple linear relationship between the covariate and the response for each treatment level is assumed.

Suppose there are t levels of a treatment factor, with each level having s_i observations; and there are r_{ij} replicates for each covariate value for treatment level i for $i = 1, 2, \ldots, t$ and $j = 1, 2, \ldots, n_i$ where n_i is the number of distinct covariate values at treatment level i. Let n be the total number of observations in the entire data set, then $s_i = \sum_{j=1}^{n_i} r_{ij}$ and $n = \sum_{i=1}^{t} s_i$. ANCOVA fits a straight line to each treatment level. These lines can be modeled as

$$Y_{ijk} = \mu_i + \beta_i X_{ij} + e_{ijk}, \tag{1}$$

where X_{ij} is the jth value of the covariate for treatment level i, Y_{ijk} is the kth response at X_{ij}, μ_i is the intercept and β_i is the slope for treatment level i, and e_{ijk} is the random error term for $i = 1, 2, \ldots, t$, $j = 1, 2, \ldots, n_i$, and $k = 1, 2, \ldots, r_{ij}$. The assumptions for the ANCOVA model in Eq. (1) are: (a) the X_{ij} (covariate) values are assumed to be fixed (i.e., X_{ij} is not a random variable); (b) $e_{ijk} \overset{iid}{\sim} N(0, \sigma_e^2)$ for all treatments where $\overset{iid}{\sim}$ stands for "independently identically distributed as". This implies Y_{ijk} are independent of each other and $Y_{ijk} \sim N(\mu_i + \beta_i X_{ij}, \sigma_e^2)$; (c) the covariate and the treatment factors are independent. Then the straight line fitted by ANCOVA to each treatment can be written as $\widehat{Y}_{ij} = \hat{\mu}_i + \hat{\beta}_i X_{ij}$, where $\widehat{Y}_{ij}$ is the predicted response for treatment i at X_{ij}, $\hat{\mu}_i$ is the estimated intercept, and $\hat{\beta}_i$ is the estimated slope for treatment i.

In the analysis, first we test

(i) $H_o : \beta_1 = \beta_2 = \cdots = \beta_t = 0$ (all slopes are equal to zero).

If H_o is not rejected, then the covariate is not necessary in the model. Then a regular one-way ANOVA can be performed to test the equality of treatment means. If H_o in (i) is rejected, then we test

$$\text{(ii)} \quad H_o : \beta_1 = \beta_2 = \cdots = \beta_t \quad \text{(the slopes are equal).}$$

If H_o in (ii) is not rejected, then the lines are parallel, otherwise they are non parallel (Milliken and Johnson, 2002).

The parallel lines model is given by

$$Y_{ijk} = \mu_i + \beta X_{ij} + e_{ijk}, \tag{2}$$

where β is the common slope for all treatment levels. With this model, testing the equality of the intercepts, i.e., $H_o : \mu_1 = \mu_2 = \cdots = \mu_t$, is equivalent to testing the equality of treatment means at any value of the covariate. For the non parallel lines case, the comparison of treatments may give different results at different values of the covariate.

2.2. *Analysis of Covariate-Adjusted Residuals*

First an overall regression line is fitted to the entire data set as:

$$\widehat{Y}_{ij} = \hat{\mu} + \hat{\beta}^* X_{ij}, \quad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, 2, \ldots, n_i, \tag{3}$$

where $\hat{\mu}$ is the estimated overall intercept and $\hat{\beta}^*$ is the estimated overall slope. The residuals from this regression line are called *covariate-adjusted residuals* and are calculated as:

$$R_{ijk} = Y_{ijk} - \widehat{Y}_{ij} = Y_{ijk} - \hat{\mu} - \hat{\beta}^* X_{ij}, \quad \text{for } i = 1, 2, \ldots, t, \;\; j = 1, 2, \ldots, n_i, \text{ and } k = 1, 2, \ldots, r_{ij}, \tag{4}$$

where R_{ijk} is the kth residual of treatment level i at X_{ij}.

2.2.1. ANOVA with or without HOV on Covariate-Adjusted Residuals. In ANOVA with or without HOV procedures, the covariate-adjusted residuals in Eq. (4) are taken to be the response values, and tests of equal treatment means are performed on residual means. The means model and assumptions for the one-way ANOVA with HOV on these covariate-adjusted residuals are:

$$R_{ijk} = \rho_i + \varepsilon_{ijk}, \quad \text{for } i = 1, 2, \ldots, t, \;\; j = 1, 2, \ldots, n_i, \text{ and } k = 1, 2, \ldots, r_{ij}, \tag{5}$$

where ρ_i is the mean residual for treatment i, ε_{ijk} are the (independent) random errors such that $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. However, R_{ijk} are not independent of each other, since $\sum_{i=1}^{t} \sum_{j=1}^{n_i} \sum_{k=1}^{r_{ij}} R_{ijk} = 0$, which also implies that the overall mean of the residuals is zero.

For the non parallel lines model in Eq. (1), the residuals in Eq. (4) will take the form:

$$R_{ijk} = Y_{ijk} - \widehat{Y}_{ij} = \mu_i + \beta_i X_{ij} + e_{ijk} - \left(\hat{\mu} + \hat{\beta}^* X_{ij}\right) = (\mu_i - \hat{\mu}) + \left(\beta_i - \hat{\beta}^*\right) X_{ij} + e_{ijk}. \tag{6}$$

Hence, the influence of the covariate will be removed if and only if (iff)

$$\hat{\beta}^* = \beta_i \quad \text{for all } i = 1, 2, \ldots, t, \tag{7}$$

that is, iff the treatment-specific lines in Eq. (1) and the overall regression in Eq. (3) are parallel. Notice that the residuals from the treatment-specific models in Eq. (1) cannot be used as response values in an ANOVA with HOV, because treatment sums of squares of such residuals are zero (Ceyhan, 2000).

In ANOVA without HOV on covariate-adjusted residuals, the only difference from ANOVA with HOV is that ε_{ijk} are the (independent) random errors such that $\varepsilon_{ijk} \sim N(0, \sigma_i^2)$. Notice that HOV is not necessarily assumed in this model.

K–W test is an extension of the Mann–Whitney U test to three or more groups; and for two groups K–W test and Mann–Whitney U test are equivalent (Siegel and Castellan, 1988). K–W test on the covariate-adjusted residuals which are obtained as in model (4) tests the equality of the residual distributions for all treatment levels. Notice that contrary to the parametric models and tests in previous sections, only the distributional equality is assumed, neither normality nor HOV.

3. Comparison of the Methods

ANOVA with or without HOV or K–W test on covariate-adjusted residuals and ANCOVA can be compared when the treatment-specific lines and the overall regression line are parallel. For two treatments, the null hypotheses tested by "ANCOVA", "ANOVA with or without HOV", and "K–W test" on covariate-adjusted residuals are H_o : "Intercepts are equal for all treatments", H_o : "Residual means are equal for all treatments", and H_o : "Residuals have the same distribution for all treatments", respectively. More formally, these null hypotheses are

$$H_o : \mu_1 = \mu_2 \quad (\text{or } \mu_1 - \mu_2 = 0) \tag{8}$$

$$H_o : \rho_1 = \rho_2 \quad (\text{or } \rho_1 - \rho_2 = 0) \tag{9}$$

and

$$H_o : F_{R_1} = F_{R_2} \quad (\text{or } R_1 \stackrel{d}{=} R_2), \tag{10}$$

respectively, where F_{R_i} is the residual distribution function for treatment i, $i = 1, 2$, and $\stackrel{d}{=}$ stands for "equal in distribution".

In Eq. (9), ρ_i can be estimated by the sample residual mean, $\overline{R}_{i..}$. Averaging the residuals in Eq. (6) for treatment i yields

$$\overline{R}_{i..} = \rho_i + \overline{\varepsilon}_{i..} = (\mu_i - \hat{\mu}) + \left(\beta_i - \hat{\beta}^*\right)\overline{X}_{i.} + \overline{e}_{i..}, \quad i = 1, 2, \tag{11}$$

where $\overline{X}_{i.}$ is the sample mean of covariate values for treatment i, $\overline{e}_{i..} = \sum_{j=1}^{n_i} \sum_{k=1}^{r_{ij}} e_{ijk}/n_i$ and $\overline{\varepsilon}_{i..} = \sum_{j=1}^{n_i} \sum_{k=1}^{r_{ij}} \varepsilon_{ijk}/n_i$, $i = 1, 2$. Under the assumptions of ANCOVA and ANOVA (with or without HOV) on covariate-adjusted residuals, taking the expectations in (11) yields

$$\mathbf{E}[\overline{R}_{i..}] = \rho_i = \mu_i + \beta_i \overline{X}_{i.} - \mu - \beta^* \overline{X}_{i.} = \mu_i - \mu + (\beta_i - \beta^*)\overline{X}_{i.}, \quad i = 1, 2, \tag{12}$$

since $\mathbf{E}[\overline{e}_{i..}] = 0$ and $\mathbf{E}[\overline{\varepsilon}_{i..}] = 0$, for $i = 1, 2$. Hence, H_o in (9) can be rewritten as H_o : $(\mu_1 - \mu_2) + (\beta_1 - \beta^*)\overline{X}_{1.} - (\beta_2 - \beta^*)\overline{X}_{2.} = 0$. Then the hypotheses in Eqs. (8) and (9) are equivalent iff

$$(\beta_1 - \beta^*)\overline{X}_{1.} = (\beta_2 - \beta^*)\overline{X}_{2.} \tag{13}$$

Using condition (7) and repeating the above argument for all pairs of treatments, the condition in (13) can be extended to more than two treatments.

Notice that the conditions that will imply (13) will also imply the equivalence of the hypotheses in (8) and (9). The overall regression slope can be estimated as

$$\hat{\beta}^* = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}}(X_{ij} - \overline{X}_{..})(Y_{ijk} - \overline{Y}_{...})}{E^*_{xx}} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}}(X_{ij} - \overline{X}_{..})Y_{ijk}}{E^*_{xx}} \tag{14}$$

where $\overline{X}_{..}$ is the overall covariate mean, $\overline{Y}_{...}$ is the overall response mean, and

$$E^*_{xx} = \sum_{i=1}^{2}\sum_{j=1}^{n_i} r_{ij}(X_{ij} - \overline{X}_{..})^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i} r_{ij}(X_{ij} - \overline{X}_{..})X_{ij}.$$

Furthermore, the treatment-specific slope used in model (1) is estimated as

$$\hat{\beta}_i = \frac{\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}}\left(X_{ij} - \overline{X}_{i.}\right)\left(Y_{ijk} - \overline{Y}_{i..}\right)}{E_{xx,i}},$$

where $E_{xx,i} = \sum_{j=1}^{n_i} r_{ij}(X_{ij} - \overline{X}_{i.})^2$, and $\overline{Y}_{i..}$ is the mean response for treatment i. Substituting $Y_{ijk} = \hat{\mu}_i + \hat{\beta}_i X_{ij} + R'_{ijk}$, $i = 1, 2$, $j = 1, 2, \ldots, n_i$, and $k = 1, 2, \ldots, r_{ij}$ in Eq. (14) where R'_{ijk} is the kth residual at X_{ij} in model (1), the estimated overall slope becomes

$$\begin{aligned}
\hat{\beta}^* &= \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}}(X_{ij} - \overline{X}_{..})(\hat{\mu}_i + \hat{\beta}_i X_{ij} + R'_{ijk})}{E^*_{xx}} \\
&= \hat{\beta}_i + \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i} r_{ij}(X_{ij} - \overline{X}_{..})\hat{\mu}_i}{E^*_{xx}} + \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}}(X_{ij} - \overline{X}_{..})R'_{ijk}}{E^*_{xx}} \\
&= \hat{\beta}_i + \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i} r_{ij}(X_{ij} - \overline{X}_{..})\hat{\mu}_i}{E^*_{xx}} + \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}} X_{ij}R'_{ijk}}{E^*_{xx}},
\end{aligned} \tag{15}$$

since $\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}} \overline{X}_{..}R'_{ijk} = 0$. As $\mathbf{E}[R'_{ijk}] = 0$, taking the expectations in (15) yields

$$\begin{aligned}
\beta^* &= \beta_i + \frac{\mu_1(\sum_{j=1}^{n_1} r_{ij}(\overline{X}_{1j} - \overline{X}_{..})) + \mu_2(\sum_{j=1}^{n_2} r_{ij}(\overline{X}_{2j} - \overline{X}_{..}))}{E^*_{xx}} \\
&= \beta_i + \frac{\mu_1 n_1(\overline{X}_{1.} - \overline{X}_{..}) + \mu_2 n_2(\overline{X}_{2.} - \overline{X}_{..})}{E^*_{xx}}
\end{aligned} \tag{16}$$

Under $H_o : \mu_1 = \mu_2$, (16) reduces to $\beta^* = \beta_i$ iff

$$\frac{n_1(\overline{X}_{1.} - \overline{X}_{..}) + n_2(\overline{X}_{2.} - \overline{X}_{..})}{E^*_{xx}} = 0 \tag{17}$$

provided that $E^*_{xx} \neq 0$. Indeed, $E^*_{xx} = 0$ will hold if and only if all X_{ij} are equal to a constant for each treatment i, in which case, $\widehat{\beta}^*$ and $\widehat{\beta}_i$ will be undefined. The condition in (17) holds if $\overline{X}_{1.} = \overline{X}_{2.} (= \overline{X}_{..})$. Recall that $H_o : \rho_1 = \rho_2$ was shown to be equivalent to $H_o : \mu_1 = \mu_2$ provided that $(\beta_1 - \beta^*)\overline{X}_{1.} = (\beta_2 - \beta^*)\overline{X}_{2.}$, which holds if $\overline{X}_{1.} = \overline{X}_{2.}$ and $\beta_1 = \beta_2$. So the null hypotheses in (8) and (9) are equivalent when the treatment-specific lines are parallel and treatment-specific means are equal which implies the condition stated in (7).

In general for t treatments, H_o in (8) can be tested by

$$\begin{aligned} F &= \frac{MSTrt}{MSE} \\ &= \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{n_i} r_{ij}\left[(\overline{Y}_{i..} - \overline{Y}_{...}) - \widehat{\beta}_i(\overline{X}_{i.} - \overline{X}_{..})\right]^2\right)/(t-1)}{\left(\sum_{i=1}^{t}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}}\left[(Y_{ijk} - \overline{Y}_{i..}) - \widehat{\beta}_i(X_{ij} - \overline{X}_{i.})\right]^2\right)/[n-(t+1)]}, \\ &\sim F(t-1, n-t-1), \end{aligned} \tag{18}$$

where $MSTrt$ is the mean square treatment for response values and MSE is the mean square error for response values.

Similarly, H_o in (9) can be tested by $F^* = MSTrt^*/MSE^*$ where $MSTrt^*$ is the mean square treatment for covariate-adjusted residuals and MSE^* is the mean square error for covariate-adjusted residuals. These mean square terms can be calculated as $MSTrt^* = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i} r_{ij}(\overline{R}_{i..} - \overline{R}_{...})^2}{(t-1)}$ and $MSE^* = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}}(R_{ijk} - \overline{R}_{i..})^2}{(n-t)}$. Using $\overline{R}_{i..} = \overline{Y}_{i..} - \hat{\mu} - \hat{\beta}^* \overline{X}_{i.}$, $i = 1, 2, \ldots, t$, and $\overline{R}_{...} = \overline{Y}_{...} - \hat{\mu} - \hat{\beta}^* \overline{X}_{..}$, we have

$$\begin{aligned} F^* &= \frac{MSTrt^*}{MSE^*} \\ &= \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{n_i} r_{ij}\left[(\overline{Y}_{i..} - \overline{Y}_{...}) - \hat{\beta}^*(\overline{X}_{i.} - \overline{X}_{..})\right]^2\right)/(t-1)}{\left(\sum_{i=1}^{t}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}}\left[(Y_{ijk} - \overline{Y}_{i..}) - \hat{\beta}^*(X_{ij} - \overline{X}_{i.})\right]^2\right)/(n-t)} \\ &\sim F\left(t-1, v^*\right). \end{aligned} \tag{19}$$

It might seem that MSE^* has $v^* = (n-t)$ degrees of freedom (df), since there are t parameters (ρ_i for $i = 1, 2, \ldots, t$) to estimate, so the test statistic in Eq. (19) is distributed as $F^* \sim F(t-1, n-t)$. However, there is one more restriction in test (9). Since $\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{r_{ij}} R_{ijk} = 0$, F^* should actually be distributed as $F^* \sim F(t-1, n-t-1)$. Atchley et al. (1976) did not suggest this adjustment in df, and Beaupre and Duvall (1998) used the method without such an adjustment. That is, in both sources $F(t-1, n-t)$ is used for inference. So, in this article df for MSE^* has been set to $(n-t)$ as in literature for comparative purposes.

For two treatments, $F \overset{d}{=} \mathcal{T}^2(n-3)$ and $F^* \overset{d}{=} \mathcal{T}^2(n-2)$ where $\mathcal{T}(n)$ is the t-distribution with n df. As $n \to \infty$, both F and F^* will converge in distribution to χ^2_1. So F and F^* will yield similar results for large n.

The above discussion is based on normality of error terms with HOV. Without HOV the df of the F-tests are calculated with Satterthwaite approximation (Kutner et al., 2004). On the other hand, K–W test requires neither normality nor HOV, but implies a more general hypothesis $H_o : F_{R_1} = F_{R_2}$, in the sense that $F_{R_1} = F_{R_2}$ would imply $\rho_1 = \rho_2$ without the normality assumption. However, the null hypothesis in Eq. (9) implicitly assumes normality.

4. Monte Carlo Simulation Analysis

4.1. *Sample Generation for Null and Alternative Models*

Without loss of generality, the slope in model (2) is arbitrarily taken to be 2 and the intercept is chosen to be 1. So the response values for the treatments are generated as

$$\text{(i)} \quad Y_{1jk} = 1 + 2X_{1j} + e_{1jk}, \quad j = 1, 2, \ldots, n_1 \text{ and } k = 1, 2, \ldots, r_{1j} \text{ for treatment 1} \tag{20}$$

and

$$\text{(ii)} \quad Y_{2jk} = (1 + 0.02q) + 2X_{2j} + e_{2jk}, \quad j = 1, 2, \ldots, n_2 \text{ and } k = 1, 2, \ldots, r_{2j} \text{ for treatment 2,} \tag{21}$$

where $e_{ijk} \overset{iid}{\sim} F_i$, where F_i is the error distribution for treatment i, $i = 1, 2$ and q is introduced to obtain separation between the parallel lines. In (20) and (21), X_{ij} is the jth generated value of the covariate in treatment i, Y_{ijk} is the response value for treatment level i at X_{ij} for $i = 1, 2$, e_{ijk} is the kth random error term. The covariate ranges, sample sizes (n_1 and n_2), error distributions (F_1 and F_2) for the two treatments, and the number of replicates (reps) at each value of X_{ij} are summarized in Table 1. In the context of model (2) the common slope is $\beta = 2$, and $\mu_1 = 1$ and $\mu_2 = (1 + 0.02q)$ are the intercepts for treatment levels 1 and 2, respectively.

As q increases the treatment-specific response means become farther apart at each covariate value and the power of the tests is expected to increase. q is incremented from 1 to m_u in case-u, for $u = 1, 2, \ldots, 16$ (Table 1) where m_u is estimated by the standard errors of the intercepts of the treatment-specific regression lines. In the simulation no further values of q are chosen when the power is expected to approach 1.00 that occurs when the intercepts are approximately 2.5 standard errors apart, as determined by equating the intercept difference, $0.02q = 2.5s_{\hat{\mu}_i}$, with q replaced by m_u. A pilot sample of size 6,000 is generated ($q = 0, 1, 2, 3, 4, 5$ with 1,000 samples at each q), and maximum of the standard errors of the intercepts is recorded. Then $m_u \cong 2.5 \max_i(s_{\hat{\mu}_i})/0.02$ for $i = 1, 2$ in case u.

Henceforth, all cases labeled with "a" have one replicate and all cases labeled with "b" have two replicates per covariate value, henceforth. For example, in case 1a the most general case is simulated with *iid* $N(0, 1)$ error variances, and 20 uniformly randomly generated covariate values in the interval $(0, 10)$ for both treatments. In case 1b, the data is generated as in case 1a with two replicates per covariate value.

In cases 1, 5–8, 9, and 12–16, error variances are homogeneous; in cases 1 and 5–8 error terms are generated as *iid* $N(0, 1)$. In case 9, error terms are generated as *iid* $\mathcal{U}(-\sqrt{3}, \sqrt{3})$; in case 12, error terms are *iid* $DW(0, 1, 3)$, double-Weibull

Table 1

The simulated cases for the comparison of ANCOVA and methods on covariate-adjusted residuals. e_{ijk}: error term; $\overset{ind}{\sim}$: independently distributed as; n_i: sample size for treatment level $i = 1, 2$. $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2; $\mathcal{U}(a, b)$ is the uniform distribution with support (a, b); $DW(a, b, c)$ is the double Weibull distribution with location parameter a, scale parameter b, and shape parameter c; $\beta(a, b)$ is the Beta distribution with shape parameters a and b; χ^2_2 is the chi-square distribution with 2 df; $LN(a, b)$ is the log-normal distribution with location parameter a and scale parameter b

Case	Error term		Sample sizes		Ranges of covariate for	
	$e_{1jk} \overset{ind}{\sim}$	$e_{2jk} \overset{ind}{\sim}$	n_1	n_2	Treatment 1	Treatment 2
1	$N(0, 1)$	$N(0, 1)$	20	20	(0, 10)	(0, 10)
2	$N(0, 1)$	$N(0, 6)$	20	20	(0, 10)	(0, 10)
3	$N(0, 1)$	$N(0, \sqrt{x})$	20	20	(0, 10)	(0, 10)
4	$N(0, \sqrt{x})$	$N(0, \sqrt{x})$	20	20	(0, 10)	(0, 10)
5	$N(0, 1)$	$N(0, 1)$	28	12	(0, 10)	(0, 10)
6	$N(0, 1)$	$N(0, 1)$	20	20	(0, 6)	(4, 10)
7	$N(0, 1)$	$N(0, 1)$	20	20	$(0, 3) \cup (7, 10)$	(4, 10)
8	$N(0, 1)$	$N(0, 1)$	20	20	$(0, 4) \cup (6, 10)$	(3, 7)
9	$\mathcal{U}(-\sqrt{3}, \sqrt{3})$	$\mathcal{U}(-\sqrt{3}, \sqrt{3})$	20	20	(0, 10)	(0, 10)
10	$\mathcal{U}(-\sqrt{3}, \sqrt{3})$	$\mathcal{U}(-2\sqrt{3}, 2\sqrt{3})$	20	20	(0, 10)	(0, 10)
11	$\mathcal{U}(-\sqrt{3}, \sqrt{3})$	$\mathcal{U}(-\sqrt{x}, \sqrt{x})$	20	20	(0, 10)	(0, 10)
12	$DW(0, 1, 3)$	$DW(0, 1, 3)$	20	20	(0, 10)	(0, 10)
13	$\sqrt{48}(\beta(6, 2) - 3/4)$	$\sqrt{48}(\beta(6, 2) - 3/4)$	20	20	(0, 10)	(0, 10)
14	$\chi^2_2 - 2$	$\chi^2_2 - 2$	20	20	(0, 10)	(0, 10)
15	$LN(0, 1) - e^{1/2}$	$LN(0, 1) - e^{1/2}$	20	20	(0, 10)	(0, 10)
16	$N(0, 2)$	$\chi^2_2 - 2$	20	20	(0, 10)	(0, 10)

distribution with location parameter 0, scale parameter 1, and shape parameter 3 whose probability density function (pdf) is $f(x) = (3/2)x^2 \exp\left(-|x|^3\right)$ for all x; in case 13, error terms are $iid\, \sqrt{48}(\beta(6, 2) - 3/4)$ where $\beta(6, 2)$ is the Beta distribution with shape parameters 6 and 2 whose pdf is $f(x) = 42x^5(1 - x)\mathbf{I}(0 < x < 1)$ where $\mathbf{I}(\cdot)$ is the indicator function; in case 14, error terms are $iid\, \chi^2_2 - 2$ where χ^2_2 is the chi-square distribution with 2 df; in case 15, error terms are $iid\, LN(0, 1) - e^{1/2}$ where $LN(0, 1)$ is the log-normal distribution with location parameter 0 and scale parameter 1 whose pdf is $f(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\log x)^2\right)\mathbf{I}(x > 0)$, and in case 16, error terms are $iid\, N(0, 2)$ for treatment 1 and iid $\chi^2_2 - 2$ for treatment 2.

In cases 2–4, heterogeneity of variances for normal error terms is introduced either by unequal but constant variances (case 2), unequal but a combination of constant and x-dependent variances (case 3), or equal and x-dependent variances (case 4). In case 10, error terms are $iid\, \mathcal{U}(-\sqrt{3}, \sqrt{3})$ for treatment 1 and $iid\, \mathcal{U}(-2\sqrt{3}, 2\sqrt{3})$ for treatment 2; in case 11, error terms are $iid\, \mathcal{U}(-\sqrt{3}, \sqrt{3})$ treatment 1 and $iid\, \mathcal{U}(-\sqrt{x}, \sqrt{x})$ for treatment 2.

The x-dependence of variances is a realistic but not a general case. For example, Beaupre and Duvall (1998) who explored the differences in metabolism (O_2 consumption) of the Western diamondback rattlesnakes with respect to their sex, the O_2 consumption was measured for males, non reproductive females, and

vitellogenic females. To remove the influence of body mass which was deemed as a covariate on O_2 consumption, ANOVA with HOV on covariate-adjusted residuals was performed. In their study, the variances of O_2 consumption for sexual groups have a positive correlation with body mass. In this study, $\sqrt{x}$ is taken as the variance term to simulate such a case. Heterogeneity of variances conditions violate one of the assumptions for ANCOVA and ANOVA with HOV on covariate-adjusted residuals, and are simulated in order to evaluate the sensitivity of the methods to such violations. In case 5, different sample sizes are taken from that of other cases to see the influence of unequal sample sizes.

The error terms are generated from normal distributions in cases 1–8, non normal distributions in cases 9–15. The distribution of the error variances are symmetric around 0 in cases 9–12, but not in cases 13–15. Notice that cases 13–15 are normalized to have zero mean, and furthermore case 13 is scaled to have unit variance. The influence of non normality and asymmetry of the distributions are investigated in these cases. In case 16, the influence of distributional differences (normal vs asymmetric non normal) in the error term is investigated.

In cases 1–5 and 9–16, covariates are uniformly randomly generated, without loss of generality, in $(0, 10)$, hence $\overline{X}_{1.} \approx \overline{X}_{2.}$ is expected to hold. In these cases the influence of replications (or magnitude of equal sample sizes), heterogeneity of variances, and non normality of the variances on the methods are investigated. Cases 6–8 address the issue of clustering which might result naturally in a data set. Clustering occurs if the treatments have distinct or partially overlapping ranges of covariates. Extrapolation occurs if the clusters are distinct or the mean of the covariate is not within the covariate clusters for at least one treatment. In case 6 there is a mild overlap of the covariate clusters for treatments 1 and 2, such that covariates are uniformly randomly generated within $(0, 6)$ for treatment 1, and $(4, 10)$ for treatment 2, so $\overline{X}_{1.}$ and $\overline{X}_{2.}$ are expected to be different. In fact, this case is expected to contain the largest difference between $\overline{X}_{1.}$ and $\overline{X}_{2.}$. In case 7, treatment 1 has two clusters, such that each treatment 1 covariate is randomly assigned to either $(0, 3)$ or $(7, 10)$ first, then the covariate is uniformly randomly generated in that interval. Treatment 2 covariates are generated uniformly within the interval of $(4, 10)$. Note that $\overline{X}_{1.}$ and $\overline{X}_{2.}$ are expected to be very different, but not as much as case 6. See Ceyhan and Goad (2009) for realizations of cases 6 and 7. Notice that the second cluster of treatment 1 is completely inside the covariate range of treatment 2. These choices of clusters are inspired by the research of Beaupre and Duvall (1998). In case 8, treatment 1 has two clusters, each treatment 1 covariate is uniformly randomly generated in the randomly selected interval of either $(0, 4)$ or $(6, 10)$. Treatment 2 covariates are uniformly randomly generated in the interval $(3, 7)$. Hence $\overline{X}_{1.}$ and $\overline{X}_{2.}$ are expected to be similar. Notice that treatment 2 cluster is in the middle of the treatment 1 clusters with mild overlaps.

4.2. *Monte Carlo Simulation Results*

4.2.1. Empirical Size Comparisons. In the simulation process, for each case, $N_{mc} = 10{,}000$ samples are generated with $q = 0$ using the relationships in (20) and (21). Out of these 10,000 samples the number of significant treatment differences detected by the methods is recorded and is used to estimate the empirical sizes. The nominal significance level used in all these tests is $\alpha = 0.05$. The number of differences detected concurrently by each pair of methods is also recorded to estimate the

proportion of agreement between each pair of methods. Using the asymptotic normality of proportions, the 95% confidence intervals are constructed for empirical sizes of the methods (not presented) to see whether they contain the nominal significance level of 0.05 and the 95% confidence interval for the difference in the proportions (not presented either) to check whether the sizes are significantly different from each other.

The empirical size estimates in cases 1a–16a and 1b–2b are presented in Table 2. Observe that ANCOVA method is liberal in case 2a and conservative in cases 14a and 15a, and has the desired nominal level 0.05 for the other cases. The liberalness in case 2a weakens as the number of replicates is doubled (see case 2b). ANOVA with or without HOV are liberal in cases 1a, 2a, and 3a, and conservative in cases 6a–8a, and 14a–15a and have the desired nominal level for the other cases. However, the liberalness of the tests weakens in cases 1a–3a, as the number of replicates is doubled. K–W test is liberal in cases 1a–3a, 10a, 11a, and 16a, and conservative in cases 6a, 7a, and 14a, and has the desired nominal level for the other cases.

Table 2

The empirical sizes and their comparisons for ANCOVA and methods on covariate-adjusted residuals for the 16 cases listed in Table 1 based on 10,000 Monte Carlo samples: $\hat{\alpha}_i$: empirical size of method i; (i, j): empirical size comparison of method i vs. method j for $i, j = 1, 2, 3, 4$ with $i \neq j$ where method $i = 1$ is for ANCOVA, $i = 2$ and $i = 3$ are for ANOVA with and without HOV on covariate-adjusted residuals, respectively, $i = 4$ is for K–W test covariate-adjusted residuals. $^{\ell}(^{c})$: Empirical size is significantly larger (smaller) than 0.05; i.e., method is liberal (conservative). $\approx$: Empirical sizes are not significantly different from each other; i.e., methods do not differ in size. $<$ ($>$): Empirical size of the first method is significantly smaller (larger) than the second

	Empirical sizes				Size comparison					
Case	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	(1, 2)	(1, 3)	(1, 4)	(2, 3)	(2, 4)	(3, 4)
1a	.0531	$.0541^{\ell}$	$.0540^{\ell}$	.0532	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
1b	.0507	.0493	.0493	.0510	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
2a	$.0581^{\ell}$	$.0576^{\ell}$	$.0546^{\ell}$	$.0612^{\ell}$	$\approx$	$\approx$	$<$	$\approx$	$<$	$<$
2b	.0531	.0515	.0493	$.0630^{\ell}$	$\approx$	$\approx$	$<$	$\approx$	$<$	$<$
3a	$.0606^{\ell}$	$.0602^{\ell}$	$.0567^{\ell}$	$.0693^{\ell}$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$<$
4a	.0523	.0525	.0519	.0511	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
5a	.0490	.0496	.0499	.0502	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
6a	$.0556^{\ell}$	$.0024^{c}$	$.0024^{c}$	$.0033^{c}$	$>$	$>$	$>$	$\approx$	$\approx$	$\approx$
7a	.0465	$.0339^{c}$	$.0337^{c}$	$.0332^{c}$	$>$	$>$	$>$	$\approx$	$\approx$	$\approx$
8a	.0474	$.0437^{c}$	$.0433^{c}$	$.0440^{c}$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
9a	.0485	.0489	.0484	.0488	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
10a	.0508	.0505	.0490	$.0595^{\ell}$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$<$
11a	.0522	.0515	.0511	$.0576^{\ell}$	$\approx$	$\approx$	$<$	$\approx$	$<$	$<$
12a	.0490	.0494	.0492	.0491	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
13a	.0486	.0481	.0480	.0473	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
14a	$.0442^{c}$	$.0435^{c}$	$.0417^{c}$	$.0451^{c}$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$	$\approx$
15a	$.0383^{c}$	$.0386^{c}$	$.0357^{c}$	.0521	$\approx$	$\approx$	$<$	$\approx$	$<$	$<$
16a	.0510	.0514	.0502	$.0701^{\ell}$	$\approx$	$\approx$	$<$	$\approx$	$<$	$<$

Liberalness of the test in case 1a weakens as the number of replicates is doubled (see case 1b). Notice that the ANCOVA method has the desired size when the error term is normally distributed or has a symmetric distribution, tends to be slightly liberal when HOV is violated, and is conservative when error distribution is non normal and not symmetric. On the other hand, ANOVA with or without HOV have about the same size for all cases. Both methods have the desired size when error terms are normally distributed, or have symmetric distribution, and the covariates have similar means. When error terms are normal without HOV, both methods are liberal with ANOVA without HOV being less liberal. When error terms are non normal with asymmetric distributions, both methods tend to be slightly conservative. But, when the covariate means are extremely different, both methods are extremely conservative (see cases 6 and 7). See Fig. 1 for the empirical size estimates for ANCOVA and ANOVA with HOV on covariate-adjusted residuals as a function of distance between treatment-specific means. As the distance between treatment-specific means increase the empirical size for the ANOVA with HOV on covariate-adjusted residuals decreases, while the empirical size for ANCOVA is stable about the desired nominal level 0.05. K–W test has the desired level when error terms have symmetric and identical distributions, is liberal when errors have the same distribution without HOV and different distributions, and is conservative when errors have asymmetric distributions provided the covariates have similar means. But when the covariate means are very different, KW test is also extremely conservative (see cases 6 and 7).

Moreover, when the covariates have similar means, ANCOVA and ANOVA (with or without HOV) methods have similar empirical sizes. These three methods have similar sizes as K–W test when the error distributions have HOV. Without HOV, K–W test has significantly larger empirical size. When the covariate means are

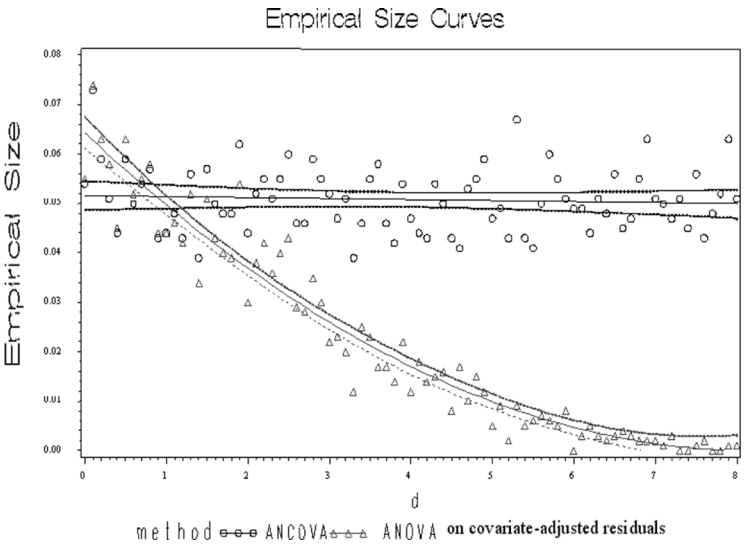

Figure 1. Empirical sizes for ANCOVA and ANOVA (with HOV) on covariate-adjusted residuals versus the distance between the treatment-specific means, $d = \overline{X}_{1.} - \overline{X}_{2.}$, with the corresponding 95% confidence bands.

considerably different, ANCOVA method has significantly larger size than others. ANOVA with or without HOV methods have similar empirical sizes for all cases.

As seen in Table 3, the proportion of agreement between the empirical size estimates are usually not significantly different from the minimum of each pair of tests for ANCOVA and ANOVA with or without HOV, but the proportion of agreement is usually significantly smaller for the cases in which K–W test is compared with others. Therefore, ANCOVA and ANOVA with or without HOV have the same null hypothesis, with similar acceptance/rejection regions, while K–W test has a different null hypothesis hence different acceptance/rejection regions. Both ANOVA methods have the same null hypothesis, and have similar acceptance/rejection regions for this simulation study.

4.2.2. Empirical Power Comparisons. Empirical power of the tests are estimated under the alternatives of treatment differences (i.e., $0.02q$ for $q \geq 1$). Then the power curves are plotted as a function of intercept difference (i.e., $0.02q$) in Figs. 2 and 3. In Fig. 3, we present cases 2a, 4a, 6a, 10a, 12a, and 15a only. For other cases, see Ceyhan (2000) and Ceyhan and Goad (2009).

Table 3

The proportion of agreement values for pairs of methods in rejecting the null hypothesis for the 16 cases listed in Table 1 based on 10,000 Monte Carlo samples: $\hat{\alpha}_{i,j}$: proportion of agreement between method i and method j in rejecting the null hypothesis for $i, j = 1, 2, 3, 4$ with $i \neq j$ where method labeling is as in Table 2. ^{n}: Proportion of agreement, $\hat{\alpha}_{i,j}$, is not significantly different from the minimum of $\hat{\alpha}_i$ and $\hat{\alpha}_j$. ^{s}: Proportion of agreement, $\hat{\alpha}_{i,j}$, is significantly smaller than the minimum of $\hat{\alpha}_i$ and $\hat{\alpha}_j$

	Proportion of agreement					
Case	$\hat{\alpha}_{1,2}$	$\hat{\alpha}_{1,3}$	$\hat{\alpha}_{1,4}$	$\hat{\alpha}_{2,3}$	$\hat{\alpha}_{2,4}$	$\hat{\alpha}_{3,4}$
1a	$.0520^n$	$.0519^n$	$.0429^s$	$.0540^n$	$.0432^s$	$.0431^s$
1b	$.0490^n$	$.0490^n$	$.0415^s$	$.0493^n$	$.0413^s$	$.0413^s$
2a	$.0560^n$	$.0545^n$	$.0419^s$	$.0546^n$	$.0415^s$	$.0405^s$
2b	$.0513^n$	$.0493^n$	$.0383^s$	$.0493^n$	$.0377^s$	$.0369^s$
3a	$.0581^n$	$.0565^n$	$.0468^s$	$.0567^n$	$.0469^s$	$.0453^s$
4a	$.0507^n$	$.0505^n$	$.0382^s$	$.0519^n$	$.0380^s$	$.0378^s$
5a	$.0473^n$	$.0389^s$	$.0382^s$	$.0396^s$	$.0392^s$	$.0388^s$
6a	$.0024^n$	$.0024^n$	$.0033^n$	$.0024^n$	$.0015^n$	$.0015^n$
7a	$.0338^n$	$.0336^n$	$.0286^s$	$.0337^n$	$.0260^s$	$.0260^s$
8a	$.0426^n$	$.0423^n$	$.0346^s$	$.0433^n$	$.0340^s$	$.0338^s$
9a	$.0475^n$	$.0473^n$	$.0417^s$	$.0484^n$	$.0422^s$	$.0420^s$
10a	$.0498^n$	$.0488^n$	$.0420^s$	$.0490^n$	$.0421^s$	$.0412^s$
11a	$.0507^n$	$.0504^n$	$.0456^s$	$.0511^n$	$.0457^s$	$.0454^s$
12a	$.0477^n$	$.0476^n$	$.0378^s$	$.0492^n$	$.0383^s$	$.0383^s$
13a	$.0476^n$	$.0476^n$	$.0369^s$	$.0480^n$	$.0371^s$	$.0371^s$
14a	$.0425^n$	$.0412^n$	$.0274^s$	$.0417^n$	$.0275^s$	$.0272^s$
15a	$.0367^n$	$.0355^n$	$.0253^s$	$.0357^n$	$.0252^s$	$.0246^s$
16a	$.0497^n$	$.0493^n$	$.0394^s$	$.0502^n$	$.0392^s$	$.0389^s$

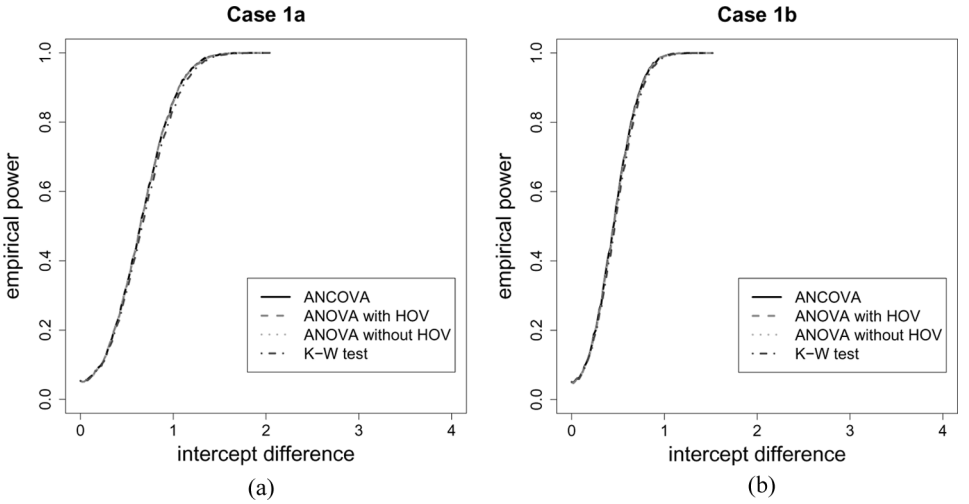

Figure 2. Empirical power estimates vs. intercept difference for cases 1a and 1b.

The first intercept difference value at which the power reaches 1 are denoted as κ and are provided in Table 4 for all cases. Observe also that power curves are steeper when error variances are smaller. In cases 1, 9–11, and 16 (of which only case 10a is presented in Fig. 3), the power estimates for ANCOVA and ANOVA methods are similar but all are larger than the K–W test power estimates. In these cases, except in cases 11 and 16, the error distributions are identical for both treatment levels, and are all symmetric; furthermore, uniform distribution approaching asymptotic normality considerably fast seems to satisfy the assumptions of the parametric tests. In cases 3, 4, 14, and 15 (of which only cases 4a and 15a are presented in Fig. 3), power estimates for ANCOVA and ANOVA methods are similar but all are smaller than those of the K–W test. In these cases, either HOV is violated as in cases 3 and 4, or normality is violated as in cases 14 and 15 with the error distribution being asymmetric. Since K–W test is non parametric, it is robust to non normality, and since it tests distributional equality, it is more sensitive to departures from HOV in normal cases. In case 5, power estimates of ANCOVA and ANOVA with HOV are similar, with both being larger than that of ANOVA without HOV whose power estimate is larger than that of K–W test. In this case, the sample sizes for the treatments are different with everything else being same. In cases 6–8 (of which only case 6a is presented in Fig. 3), the power estimate of ANCOVA method is significantly larger than those of the ANOVA methods whose empirical sizes are larger than that of K–W test. In these cases, the covariates are clustered with very different treatment-specific means in cases 6 and 7, and similar means in case 8. In cases 2 and 12 (both of which are presented in Fig. 3), for smaller values of intercept difference (i.e., between 0 to 0.5 in case 2 and 0 to 0.8 in case 12), ANCOVA and ANOVA methods have similar power with all having a smaller power than that of K–W test, while for larger values of the intercept difference (i.e., between 0.5 to 4 in case 2 and 0.8 to 2 in case 12), the order is reversed for the power estimates. In case 2, error terms have different but constant variances, and in case 12, error terms are non normal but symmetric.

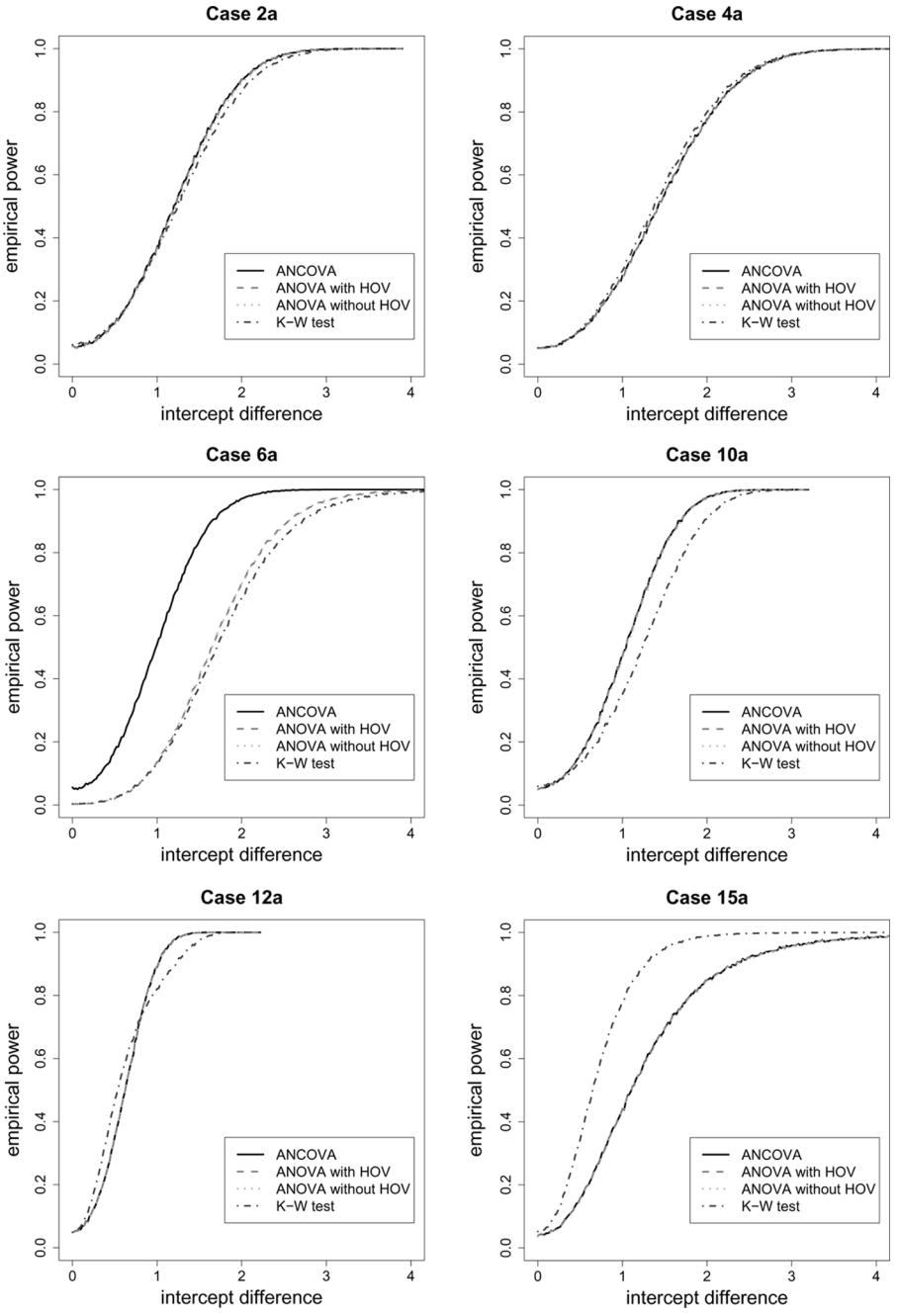

Figure 3. Empirical power estimates versus intercept difference for cases 2a, 4a, 6a, 10a, 12a, and 15a.

Table 4

The intercept difference values at which the power estimates reach 1 for the 16 cases listed in Table 1 based on 10,000 Monte Carlo samples: κ_i = intercept difference value at which power estimate of method i reaches 1 for the first time for $i = 1, 2, 3, 4$ where method labeling is as in Table 2

	Cases							
	1a; 1b	2a; 2b	3a; 3b	4a; 4b	5a; 5b	6a; 6b	7a; 7b	8a; 8b
κ_1	1.82; 1.30	3.58; 2.38	3.34; 2.38	4.06; 3.06	1.98; 1.42	2.96; 2.20	2.02; 1.40	1.98; 1.32
κ_2	1.82; 1.34	3.58; 2.50	3.34; 2.38	4.40; 3.06	2.06; 1.42	5.36; 3.30	2.28; 1.58	2.02; 1.40
κ_3	1.82; 1.34	3.58; 2.50	3.34; 2.38	4.40; 3.06	2.06; 1.42	5.36; 3.30	2.28; 1.58	2.02; 1.40
κ_4	1.90; 1.36	3.80; 2.60	3.36; 2.38	4.38; 2.92	2.06; 1.46	7.04; 3.58	2.70; 1.76	2.04; 1.46
	9a; 9b	10a; 10b	11a; 11b	12a; 12b	13a; 13b	14a; 14b	15a; 15b	16a; 16b
κ_1	1.80; 1.30	2.74; 1.98	2.06; 1.44	1.74; 1.18	1.86; 1.22	4.46; 2.78	9.86; 5.58	338; 2.34
κ_2	1.80; 1.30	2.74; 1.98	2.06; 1.44	1.74; 1.26	1.86; 1.22	4.46; 2.78	9.86; 5.58	3.42; 2.34
κ_3	1.80; 1.30	2.74; 1.98	2.06; 1.44	1.74; 1.26	1.86; 1.22	4.46; 2.78	9.86; 5.58	3.42; 2.34
κ_4	2.02; 1.52	3.20; 2.34	2.34; 1.72	2.02; 1.60	1.98; 1.32	4.10; 2.26	3.66; 1.90	3.62; 2.64

5. Discussion and Conclusions

In this article, we discuss various methods to remove the covariate influence on a response variable when testing for differences between treatment levels. The methods considered are the usual ANCOVA method and the analysis of covariate-adjusted residuals using ANOVA with or without homogeneity of variances (HOV) and Kruskal-Wallis (K–W) test. The covariate-adjusted residuals are obtained from the fitted overall regression line to the entire data set (ignoring the treatment levels). For covariate-adjusted residuals to be appropriate for removing the covariate influence, the treatment-specific lines and the overall regression line should be parallel. On the other hand, ANCOVA can be used to test the equality of treatment means at specific values of the covariate. Furthermore, the use of ANCOVA is extended to the non parallel treatment-specific lines also (Kowalski et al., 1994).

The Monte Carlo simulations indicate that when the covariates have similar means and have similar distributions (with or without HOV), ANCOVA, ANOVA with or without HOV methods have similar empirical sizes; and K–W test is sensitive to distributional differences, since the null hypotheses for the first three tests are about same while it is more general for K–W test. When the treatment-specific lines are parallel, treatment-specific covariate ranges and covariate distributions are similar. ANCOVA and ANOVA with or without HOV on covariate-adjusted residuals give similar results if error variances have symmetric distributions with or without HOV and sample sizes are similar for treatments; give similar results if error variances are homogeneous and sample sizes are different but large for treatments. In these situations, parametric tests are more powerful than K–W test. The methods give similar results but are liberal if error variances are heterogeneous with different functional forms for treatments. In these cases, usually K–W test has better performance.

When the treatment-specific lines are parallel, but treatment-specific covariate ranges are different; i.e., there exist clustering of the covariate relative to the treatment factors, ANCOVA and ANOVA on covariate-adjusted residuals yield similar results if treatment-specific covariate means are similar, very different results if treatment-specific covariate means are different since overall regression line will not be parallel to the treatment-specific lines. In such a case, methods on covariate-adjusted residuals tend to be extremely conservative whereas the size of ANCOVA F test is about the desired nominal level. Moreover, ANCOVA is much more powerful than ANOVA on covariate-adjusted residuals in these cases. The power of ANOVA on covariate-adjusted residuals gets closer to that of ANCOVA, as the difference between the treatment-specific covariate means gets smaller. However, in the case of clustering of covariates relative to the treatments, one should also exercise extra caution due to the extrapolation problem. Moreover in practice, such clustering is suggestive of an ignored grouping factor as in blocking. The discussed methods are meaningful only within the overlap of the clusters or in the close vicinity of them. However, when there are clusters for the groups in terms of the covariate, it is very likely that covariate and the group factors are dependent, which violates an assumption for ANCOVA. If this dependence is strong, then ANCOVA method will not be appropriate. On the other hand, the residual analysis is extremely conservative which might be viewed as an advantage in order not to reach spurious and confounded conclusions in such a case.

Different treatment-specific covariate distributions within the same interval or different intervals might also cause treatment-specific covariate means to be

different. In such a case, ANCOVA should be preferred against the methods on covariate-adjusted residuals.

In conclusion, we recommend the following strategy for the use of the above methods: (i) First, one should check the significance of the effect of the covariates for each treatment, i.e., test H_o^i : "all treatment-specific slopes are equal to zero". If H_o^i is not rejected, then the usual (one-way) ANOVA or K–W test can be used. (ii) If H_o^i is rejected, the covariate effect is significant for at least one treatment factor. Hence one should test H_o^{ii} : "equality of all treatment-specific slopes". If H_o^{ii} is rejected, then the covariate should be included in the analysis as an important variable and the usual regression tools can be employed. (iii) If H_o^{ii} is not rejected, check the covariate ranges. If they are similar or have a considerable intersection for treatment factors, then ANCOVA and methods on residuals are appropriate. Then one should check the underlying assumptions for the methods and then pick the best method among them. (iv) If covariate ranges are very different, then it is very likely that treatment and covariate are not independent, hence ANCOVA is not appropriate. On the other hand, the methods on residuals can be used but they are extremely conservative. In this case, one may apply some other method, e.g., MANOVA on (response, covariate) data for treatment differences.

References

Akritas, M., Arnold, S., Du, Y. (2000). Nonparametric models and methods for nonlinear analysis of covariance. *Biometrika* 87(3):507–526.

Albrecht, G. H., Gelvin, B. R., Hartman, S. E. (1993). Ratios as a size adjustment in morphometrics. *American Journal of Physical Anthropology* 91(4):441–468.

Atchley, W. R., Gaskins, C. T., Anderson, D. (1976). Statistical properties of ratios I. Empirical results. *Systematic Zoology* 25(2):137–148.

Beaupre, S. J., Duvall, D. (1998). Variation in oxygen consumption of the western diamondback rattlesnake (Crotalus atrox): Implications for sexual size dimorphism. *Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology* 168(7):497–506.

Ceyhan, E. (2000). A Comparison of Analysis of Covariance and ANOVA Methods using Covariate-Adjusted Residuals. Master's thesis, Oklahoma State University, Stillwater, OK.

Ceyhan, E., Goad, C. L. (2009). A Comparison of Analysis of Covariate-Adjusted Residuals and Analysis of Covariance. arXiv:0903.4331v2 [stat.AP]. Technical Report # KU-EC-09-4.

Garcia-Berthou, E. (2001). On the misuse of residuals in ecology: Testing regression residuals vs. the analysis of covariance. *The Journal of Animal Ecology* 70(4):708–711.

Huitema, B. E. (1980). *The Analysis of Covariance and Alternatives*. New York: John Wiley & Sons Inc.

Jakob, E. M., Marshall, S. D., Uetz, G. W. (1996). Estimating fitness: A comparison of body condition indices. *Oikos* 77(1):61–67.

Kowalski, C. J., Schneiderman, E. D., Willis, S. M. (1994). ANCOVA for non parallel slopes: The Johnson–Neyman technique. *International Journal of Bio-Medical Computing* 37(3):273–286.

Kuehl, R. O. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*. 2nd ed. Pacific Grove, CA: Brooks/Cole.

Kutner, M. H., Nachtsheim, C. J., Neter, J. (2004). *Applied Linear Regression Models*. 4th ed. Chicago: McGraw-Hill/Irwin.

Maxwell, S. E., Delaney, H. D., Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin* 95(1):136–147.

Maxwell, S. E., Delaney, H. D., Manheimer, J. M. (1985). ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational Statistics* 10(3):197–209.

Miller, G. A., Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology* 110(1):40–48.

Milliken, G., Johnson, D. E. (2002). *Analysis of Messy Data, Volume III: Analysis of Covariance*. New York: Chapman and Hall/CRC.

Ott, R. L. (1993). *An Introduction to Statistical Methods and Data Analysis*. 4th ed. Belmont, CA: Duxbury Press.

Packard, G. C., Boardman, T. J. (1988). The misuse of ratios, indices, and percentages in ecophysiological research. *Physiological Zoology* 61:1–9.

Raubenheimer, D., Simpson, S. J. (1992). Analysis of covariance: An alternative to nutritional indices. *Journal Entomologia Experimentalis et Applicata* 62(3):221–231.

Rheinheimer, D. C., Penfield, D. A. (2001). The effects of Type I error rate and power of the ANCOVA *F* test and selected alternatives under nonnormality and variance heterogeneity. *Journal of Experimental Education* 69(4):373–391.

Siegel, S., Castellan, N. J., Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York: McGraw-Hill.

Small, C. G. (1996). *The Statistical Theory of Shape*. New York: Springer-Verlag.

Tsangari, H., Akritas, M. G. (2004a). Nonparametric ANCOVA with two and three covariates. *Journal of Multivariate Analysis* 88(2):298–319.

Tsangari, H., Akritas, M. G. (2004b). Nonparametric models and methods for ancova with dependent data. *Journal of Nonparametric Statistics* 16(3–4):403–420.