Analysis of Covariance: Univariate and Multivariate Approaches

Forthcoming in the
International Encyclopedia of Education, 4th Edition

Bryan Keller & Dobrin Marchev

Teachers College, Columbia University

January 10, 2022

Abstract

In this chapter we use the general linear modeling framework to define ANCOVA and MANCOVA and compare and contrast them with their counterparts, ANOVA and MANOVA, respectively. In particular, we demonstrate how covariate adjustment leads to reduction of error variance and, given some assumptions, can remove bias due to confounding in observational studies. Data examples are used throughout to illustrate key concepts. We conclude with some general discussion about limitations and extensions.

Keywords: ANCOVA, MANCOVA, treatment effect, randomized experiment, observational study

Analysis of Covariance: Univariate and Multivariate Approaches

Forthcoming in the
International Encyclopedia of Education, 4th Edition

1 Introduction

The analysis of covariance (ANCOVA) is an inferential statistical method for analyzing experimental data that allows for the comparison of two or more group means while controlling for one or more covariates. As Fisher, who invented the procedure, explained, the analysis of covariance "combines the advantages and reconciles the requirements of the two very widely applicable procedures known as regression and analysis of variance." (1934, p. 269). ANCOVA improves upon the analysis of variance (ANOVA) in two key ways. First, by controlling for covariates that are correlated with the outcome, the ANCOVA test of treatment efficacy (i.e., difference in group means) is typically more powerful than the analogous ANOVA-based test. Second, in situations where random assignment fails to provide adequate balance on a confounding variable, or in cases where random assignment was not implemented, ANCOVA may be used to "control for" covariates by making linear adjustments to estimated group means that can reduce or remove bias due to confounding.

The use of ANCOVA to identify average treatment effects – either in the context of a well-implemented randomized experiment to reduce error variance or in a non-randomized study to control for confounding – requires assumptions, some of which are more likely to be tenable in practice than others. In modern usage, the meaning of ANCOVA is sometimes broadened to include nonlinear transformations of covariates or treatment by covariate interactions (cf. Huitema, 2011). Our focus herein, however, will be on the use of ANCOVA in the classical sense, as a method used to test for constant treatment effects assuming linear functional forms.

Multivariate analysis of variance (MANOVA) extends ANOVA to allow for simultaneous testing for treatment effects in the presence of multiple outcome variables. For correlated outcome variables, MANOVA may provide substantially better power to detect effects than the associated univariate ANOVAs. This same rationale holds when also controlling for one or more covariates via multivariate ANCOVA (MANCOVA); thus, we find reason to recommend MANCOVA for scenarios involving multiple correlated outcome variables.

The remainder of this chapter is organized as follows. In the next section we briefly define some key terms and describe the general linear F test, which is used to construct hypothesis tests in ANOVA and ANCOVA. In the third section, because central concepts in ANCOVA are most meaningful when compared against analogous concepts in ANOVA, we describe the univariate one-way ANOVA model, give a data example, and discuss pairwise comparisons of group means through contrast testing. We present the univariate ANCOVA model in Section 4, where we examine the omnibus F statistic in relation to ANOVA, give a data example, describe adjusted group means, and discuss the role of covariate balance in ANCOVA adjustment. We then follow a similar structure for the multivariate procedures, introducing MANOVA first and following with MANCOVA and a

data demonstration. We conclude with some discussion.

2 The General Linear F Test

Consider a simple linear regression model for outcome Y and predictor X,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{1}$$

where errors are independent and identically (iid) normally distributed with mean zero and constant variance; i.e., $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. When data are used to estimate the parameters of the model, the resulting function may be used to predict the mean of the outcome, Y, conditional on the values of the predictor, X. The outputs from this function are called *predictions* or *fitted values*, and the fitted value for the ith unit is denoted as $\hat{Y}_i$ as follows,

$$E[Y_i|x_i] = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are, respectively, estimates for β_0 and β_1. The *residual* for participant i, e_i, is defined as the difference in observed and predicted values; that is, $e_i = Y_i - \hat{Y}_i$; see Figure 1.

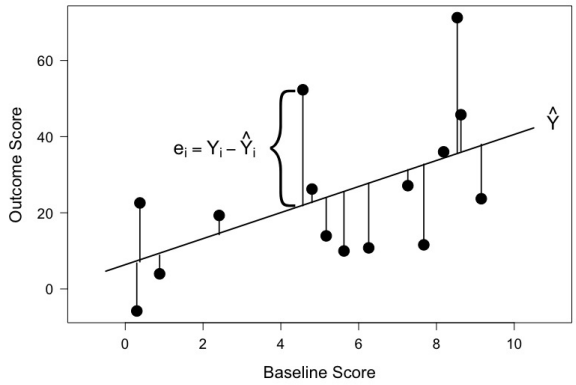

Figure 1. Scatterplot of baseline and outcome data with fitted regression line and residuals

Sums of squared residuals may be used to create measures of model fit. In particular, sums of squared residuals give a quantitative summary of variation leftover (i.e., not explained) by the model. The *general linear F test* provides a framework for constructing hypothesis tests by comparing regression models using sums of squares (SS). As an example, to test $H_0 : \beta_1 = 0$ for the simple regression model in Equation 1, a *restricted* model is specified in which the parameter to be tested (or parameters, if a joint test is desired) is fixed to the hypothesized value. Here the restricted model would be $Y_i = \beta_0 + \epsilon_i$. The general linear F statistic is defined as

$$F = \frac{(\mathrm{SS_R} - \mathrm{SS_F})/(df_R - df_F)}{\mathrm{SS_F}/df_F}, \tag{2}$$

where $\mathrm{SS_R}$ and $\mathrm{SS_F}$ are the sums of squared residuals, and df_R and df_F are degrees of

freedom, for the restricted and full models, respectively. If errors $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, then the statistic F will follow an F distribution under the null hypothesis with $df_R - df_F$ numerator and df_F denominator degrees of freedom[1].

3 Univariate ANOVA

It is helpful to use the analysis of variance (ANOVA) as a starting point from which to describe the ANCOVA model. For a single-factor experiment, the one-way ANOVA model may described as follows,

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}, \tag{3}$$

where,

- Y_{ij} is the outcome variable for participant i in group j,
- μ is the unweighted average of group means,
- α_j is the treatment effect (i.e., $\mu_j - \mu$) for the jth group, and
- ϵ_{ij} is idiosyncratic (random) error for participant i in level j.

The ϵ_{ij} are assumed to be independent $N(0, \sigma^2)$ random variables. To uniquely define model parameters, a constraint is typically imposed such that the treatment effects add to zero, $\sum_{j=1}^{a} \alpha_j = 0$[2]. The one-way ANOVA model assumes that any systematic variability in the outcome is solely due to the treatment factor groupings, and each α_j represents the jth group's deviation from the overall population mean, μ. Suppose there are a groups in total with n_j participants in the jth group. The *omnibus null hypothesis* for the one-way ANOVA is that all treatment effects are identically equal to zero; that is,

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0.$$

Table 1
Full and restricted models, parameters, and least-squares estimators for the one-way ANOVA omnibus test

ANOVA	Model	Parameters	Least-squares estimators
Full	$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$	μ	$\hat{\mu} = \bar{y}$
		$\alpha_j, \alpha_2, \ldots, \alpha_J$	$\hat{\alpha}_j = \bar{y}_j - \bar{y}$
Restricted	$Y_{ij} = \mu + \epsilon_{ij}$	μ	$\hat{\mu} = \bar{y}$

[1] See Maxwell, Delaney, and Kelley (2018) and Neter, Wasserman, and Kutner (1990) for texts that build up test statistics for ANOVA and ANCOVA via model comparisons using the general linear F test.

[2] Another possibility is to require a *weighted* sum of the α_j to be zero: $\sum_{j=1}^{a} w_j \alpha_j = 0$. In that case, μ is a weighted average of group means. For a useful discussion of weighting schemes and implications, see Neter et al. (1990, p. 554–556).

Let $\bar{y}$ and $\bar{y}_j$ be estimates of the overall mean and group j mean, respectively. The sums of squares (SS) decomposition for ANOVA is motivated by the model in Equation 3. After subtracting the mean μ from both sides, Equation 3 may be expressed as follows,

$$Y_{ij} - \mu = \alpha_j + \epsilon_{ij}, \tag{4}$$

and with sample analogues as

$$(y_{ij} - \bar{y}) = (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j). \tag{5}$$

The three terms in parentheses in Equation 5 are referred to as *deviations*. Squaring both sides, canceling out the middle term on the right hand side because $\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j) = 0$, and summing over all i and j leads to the sums of squares decomposition for one-way ANOVA; see Equation 6. Note the sum of squares "total" do not represent the total squared variation in the outcome variable because the mean has been subtracted off the left hand side of the equation. Thus, SST is typically referred to as "corrected for the mean".

$$\begin{aligned} \mathrm{SS}_{\mathrm{T;\ total\ (corrected)}} &= \mathrm{SS}_{\mathrm{B;\ between}} + \mathrm{SS}_{\mathrm{W;\ within}} \\ \sum_{j=1}^{a} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 &= \sum_{j=1}^{a} n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{a} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \end{aligned} \tag{6}$$

After substitution and some algebra, the F statistic for the one-way ANOVA omnibus null hypothesis may be written in terms of SS_B and SS_W as follows,

$$F = \frac{(\mathrm{SS}_\mathrm{R} - \mathrm{SS}_\mathrm{F})/(df_R - df_F)}{\mathrm{SS}_\mathrm{F}/df_F} = \frac{\mathrm{SS}_\mathrm{B}/(a-1)}{\mathrm{SS}_\mathrm{W}/(N-a)}, \tag{7}$$

where N is the total number of participants. If $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, the F statistic will follow an F distribution with $a - 1$ numerator and $N - a$ denominator degrees of freedom.

3.1 Data Demonstration

Consider a hypothetical experiment in which a total of $N = 15$ participants were randomly assigned to three groups (two treatments and a control) with equal frequency (i.e., $n_1 = n_2 = n_3 = 5$). All participants were measured at baseline on a variable known to be positively correlated with the outcome. The one-way ANOVA restricted model assigns predicted values to participants based on the overall mean of the outcome variable, regardless of group membership; whereas, the full model assigns a predicted value to each participant based on their group mean. The left and right panels of Figure 2 depict the data by group along with restricted and full models for the one-way ANOVA omnibus null hypothesis along with dotted lines tracing residuals that contribute to SS_R and SS_F.

ANOVA results are often displayed in tabular form as in Table 2, which reveals that $\mathrm{SS}_\mathrm{B} = 1890.1$ and $\mathrm{SS}_\mathrm{W} = 3710.0$. The mean squares of 945.1 and 309.8 represent, respectively, the numerator and denominator of the F statistic. Their quotient is 3.06 which, when compared with an F distribution with $3 - 1 = 2$ numerator and $15 - 3 = 12$ denominator *df* yields a p value of 0.084.

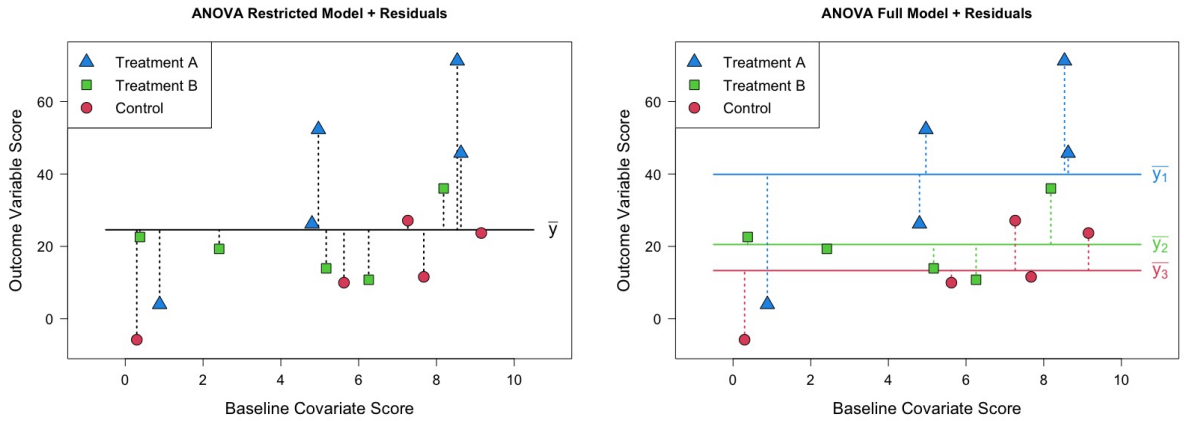

Figure 2. ANOVA restricted (left panel) and full (right panel) model fits and residuals

Table 2
ANOVA output table; SS = sum of squares; MS = mean squares; df = degrees of freedom

Source	SS	df	MS	F	p
Between	1890.1	2	945.1	3.06	0.084
Within	3710.0	12	309.8		

3.2 Comparing Group Means

A rejection of the omnibus test null hypothesis signals that at least one of the treatments had an effect, or, equivalently, that at least one pair of group means differ. Questions about which group means differ and by how much may be addressed by estimating and testing contrasts. A *contrast* is a linear combination of the group means such that the coefficients add up to 0. That is, ψ is a contrast if it may be written as follows,

$$\psi = c_1\mu_1 + c_2\mu_2 + \cdots + c_a\mu_a = \sum_{j=1}^{a} c_j\mu_j, \tag{8}$$

where $\sum_j c_j = 0$. Sample group means may be substituted to obtain an estimate,

$$\hat{\psi} = c_1\bar{y}_1 + c_2\bar{y}_2 + \cdots + c_a\bar{y}_a = \sum_{j=1}^{a} c_j\bar{y}_j, \tag{9}$$

and null hypotheses of the form $H_0 : \psi = 0$ may be tested by constructing an F statistic,

$$F_\psi = \frac{\hat{\psi}^2 \Big/ \sum_{j=1}^{a} \left(\frac{c_j}{n_j}\right)}{MS_{\text{W}}}. \tag{10}$$

The statistic F_ψ will follow an F distribution under the null hypothesis with 1 numerator and $df_F = N - a$ denominator degrees of freedom. Contrasts for pairwise comparisons of

means for treatment groups A and B and the control group, along with estimates, F statistics, and p-values, are given in Table 3. The estimated group mean differences of 19.4, 26.2, and 7.2 may be directly observed as vertical distances between prediction lines in the full ANOVA model depicted in the right panel of Figure 2.

Table 3
Table of contrasts, estimates, F statistics, and adjusted p-values for pairwise comparisons of group means

Comparison	ψ	$\hat{\psi}$	F	p
A vs. B	$1\mu_A - 1\mu_B + 0\mu_C$	19.4	3.03	0.23
A vs. Control	$1\mu_A + 0\mu_B - 1\mu_C$	26.6	5.71	0.08
B vs. Control	$0\mu_A + 1\mu_B - 1\mu_C$	7.2	0.42	0.80

Note. p-values are adjusted via Tukey's HSD for a family of three pairwise comparisons.

4 Univariate ANCOVA

The one-way ANCOVA model extends ANOVA by adding one or more numeric predictors (i.e., covariates). With a single covariate, X_{ij}, the ANCOVA model may be written as follows,

$$Y_{ij} = \mu + \alpha_j + \beta_1 X_{ij} + \epsilon_{ij}. \tag{11}$$

where β_1 is the slope that quantifies the strength of the linear relationship between the covariate, X, and the outcome, Y, and other parameters are defined as in Equation 3.

Table 4
Full and restricted models, parameters, and least-squares estimators for the one-way ANCOVA omnibus test for a model with one covariate

ANCOVA	Model	Parameters	Least-squares estimator
Full	$Y_{ij} = \mu + \alpha_j + \beta_1 X_{ij} + \epsilon_{ij}$	μ	$\hat{\mu} + \hat{\alpha}_j = \bar{y}_j - b_W \bar{x}_j$
		$\alpha_1, \alpha_2, \ldots, \alpha_a$	
		β_1	$\hat{\beta}_1 = b_W$
Restricted	$Y_{ij} = \mu + \beta_1 X_{ij} + \epsilon_{ij}$	μ	$\hat{\mu} = \bar{y} - b_T \bar{x}$
		β_1	$\hat{\beta}_1 = b_T$

The least squares estimator for coefficient β_1 under the restricted ANCOVA model, b_T, is referred to as the total or overall regression slope because it ignores group

membership:

$$b_T = \frac{\sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})(y_{ij}-\bar{y})}{\sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2}. \tag{12}$$

The least squares estimator for β_1 under the full ANCOVA model, b_W, is referred to as the pooled within-group regression slope because it estimates a single slope by pooling the numerator and denominator components that would have contributed to the within-group slopes had they been estimated separately[3]:

$$b_W = \frac{\sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)(y_{ij}-\bar{y}_j)}{\sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2}. \tag{13}$$

As with the ANOVA model, the F statistic for testing the omnibus null hypothesis, $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_j$, under the ANCOVA model may be built up by making model comparisons under the general linear F framework. In particular,

$$F = \frac{\left(\mathrm{SS_B} - \left[b_T^2 \sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 - b_W^2 \sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2\right]\right)\Big/(a-1)}{\left(\mathrm{SS_W} - b_W^2 \sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2\right)\Big/(N-a-1)} \tag{14}$$

$$= \frac{\mathrm{SS_{B(adj)}}\big/(a-1)}{\mathrm{SS_{W(adj)}}\big/(N-a-1)}, \tag{15}$$

where $\mathrm{SS_{B(adj)}}$ and $\mathrm{SS_{W(adj)}}$ are sums of squared residuals that have been *adjusted* for the linear regression of the outcome on the baseline covariate. The value of $\mathrm{SS_{B(adj)}}$ may be larger or smaller than the unadjusted (ANOVA) value of $\mathrm{SS_B}$; however, when participants are randomly assigned to groups, the value of $\mathrm{SS_{B(adj)}}$ will, on average, be identical to the unadjusted value of $\mathrm{SS_B}$.

In contrast, $\mathrm{SS_{W(adj)}}$ will always be less than or equal to $\mathrm{SS_W}$ due to the subtraction of a non-negative term, which removes within-group variability due to the linear relationship between the outcome, Y, and the covariate, X, from the residuals. The stronger the linear relationship between X and Y within groups, the greater the reduction in $\mathrm{SS_{W(adj)}}$ relative to $\mathrm{SS_W}$. If the covariate and outcome have no linear relationship in the sample, the value of b_W^2 will be zero, and the inclusion of the covariate will have no effect on $\mathrm{SS_W}$. Note that the denominator *df*, $N - a - 1$, is decreased by one relative to ANOVA because one additional parameter, β_1, is included in the ANCOVA model.

---

[3] See Wildt and Ahtola (1978) for derivations of b_T and b_W.

4.1 Data Demonstration

The left and right panels of Figure 3 depict the restricted and full models for the one-way ANCOVA omnibus null hypothesis along with dotted lines tracing residuals. Though it may be difficult to discern from the plot, the slope of the regression line in the left panel ($b_T = 3.51$) is slightly less steep than the three identical regression slopes in the right panel ($b_W = 3.71$).

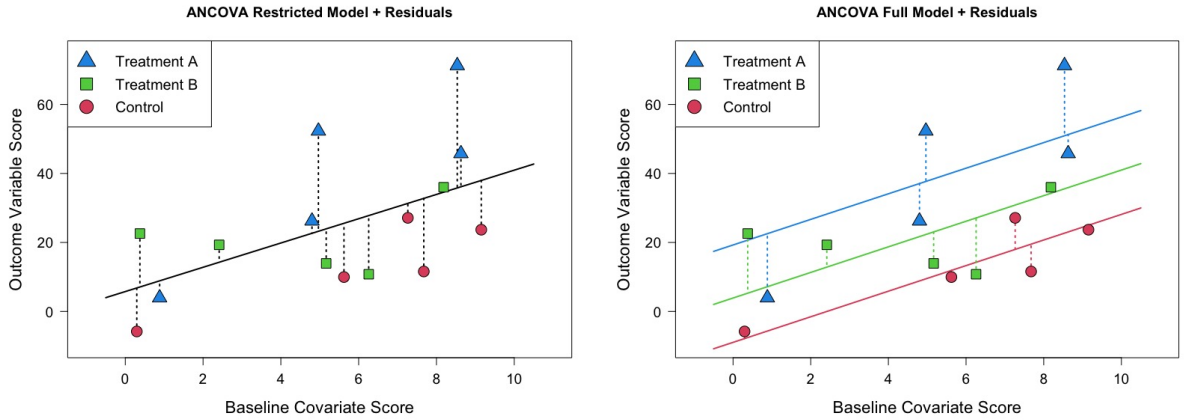

Figure 3. ANCOVA restricted (left panel) and full (right panel) model fits and residuals

ANCOVA results are displayed in Table 5. It is of interest to compare results with those of the ANOVA, shown in Table 2. Here we see that $\text{SS}_{\text{B(adj)}} = 1990.3$ as compared with $\text{SS}_{\text{B}} = 1890.1$. That this value changed by a small amount, only about 5% here, is expected because participants were randomly assigned to treatment groups and, therefore, were well-balanced on the baseline covariate. The change in within-subject variation is more drastic: $\text{SS}_{\text{W(adj)}} = 1969.4$ as compared with $\text{SS}_{\text{W}} = 3710.0$. This large reduction of approximately 50% is also expected due to the strong positive linear relationship between the covariate and the outcome, and can be seen graphically by noting that the residuals in the right panel of Figure 3 (ANCOVA full model) are substantially smaller, on average, than those in the right panel of Figure 2 (ANOVA full model).

The ANCOVA-based test of the omnibus null hypothesis ($\text{F}(2, 11) = 5.56$; $p = 0.021$) produced a larger F statistic and smaller p-value than the ANOVA-based test ($\text{F}(2, 12) = 3.06$; $p = 0.084$), largely due to the reduction in $SS_{\text{W(adj)}}$ relative to SS_{W}. That is, accounting for the linear relationship between the covariate and the outcome carved away a large portion of the variation that had previously contributed to the error variance.

4.2 Adjusted Group Means

Estimated group means under the ANCOVA full model are no longer equal to observed sample group means; instead, they are conditional on the value of the covariate (or covariates, if more than one are included).

$$\begin{aligned} E[Y_{ij}|x_{ij}] &= \hat{\mu} + \hat{\alpha}_j + b_W x_{ij} \\ &= \bar{y}_j + b_W(\bar{x}_j - x_{ij}) \end{aligned} \tag{16}$$

Table 5
ANOVA output table for the ANCOVA model; SS = sum of squares; MS = mean squares; df = degrees of freedom

Source	SS	*df*	MS	F	*p*
Between (adjusted)	1990.3	2	995.2	5.56	0.021
Covariate	1740.6	1	1740.6	9.72	0.010
Within (adjusted)	1969.4	11	179.0		

By conditioning on a single value of the covariate, ANCOVA allows us to assess what the impact of each treatment *would have been* had all the groups been centered at the same value on the covariate. For each group, $j = 1, \ldots, a$, the *adjusted group mean*, $\bar{y}'_j$, is defined as the predicted mean of the outcome score given that the covariate is set to its overall mean, $\bar{x}$.

$$\bar{y}'_j = E[Y_{ij}|x_{ij} = \bar{x}] = \bar{y}_j + b_W(\bar{x}_j - \bar{x}) \tag{17}$$

Pairwise differences in adjusted group means may be visualized as vertical distances between regression lines in the right panel of Figure 3.

Table 6
Table of contrasts for pairwise comparisons of adjusted group means

Comparison	ψ	$\hat{\psi}$	F	p-value
A vs. B	$1\mu_A - 1\mu_B + 0\mu_C$	15.4	3.22	0.216
A vs. Control	$1\mu_A + 0\mu_B - 1\mu_C$	28.2	11.08	0.017
B vs. Control	$0\mu_A + 1\mu_B - 1\mu_C$	12.9	2.21	0.335

Note. p-values are adjusted via Tukey's HSD for a family of three pairwise comparisons.

4.3 Covariate Balance, Overlap and ANCOVA Adjustment

A covariate is said to be *balanced* with respect to the treatment groups if it is identically distributed across those groups. Covariate balance is expected in well-implemented randomized experiments because random assignment renders all covariates, $\boldsymbol{X}$, statistically independent of the treatment assignment, T; that is, $\boldsymbol{X} \perp T$. In non-randomized studies or when randomization becomes complicated by post-treatment factors such as differential attrition, covariates may be systematically imbalanced[4]. When an important baseline covariate is imbalanced, adjustment via ANCOVA may cause

[4] While it is possible that a well-implemented randomized experiment may result in poor baseline covariate balance due to an unlucky randomization, this becomes less and less likely as sample sizes increase. For small to moderate samples, where chance imbalances are more likely, rerandomization (Morgan & Rubin, 2014; Zhou, Ernst, Morgan, Rubin, & Zhang, 2018) may be used to ensure a minimum level of balance is achieved.

unadjusted and adjusted group means and, hence, $SS_{B(adj)}$ and SS_B, to differ systematically. The direction of the difference depends on the severity of the imbalance and the magnitudes of b_T and b_W; see the numerator of Equation 14.

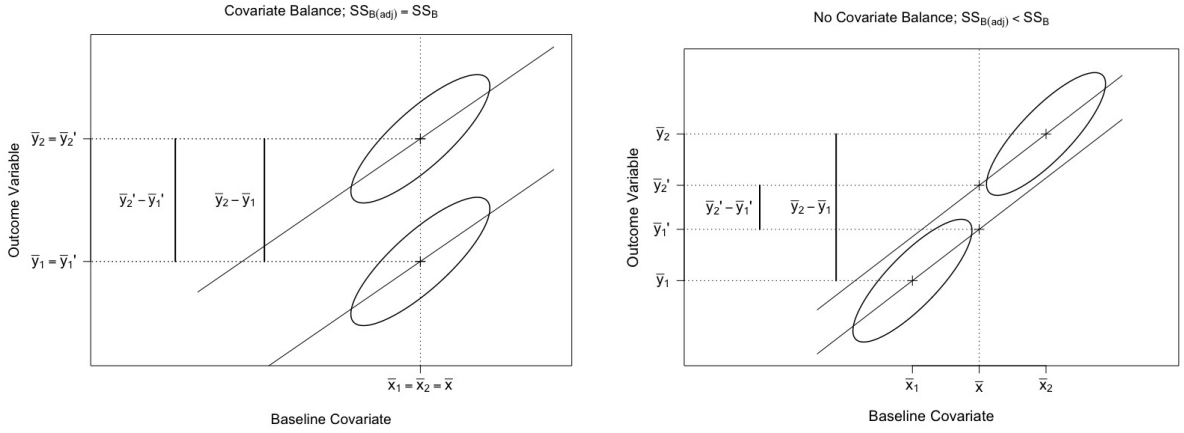

Figure 4. A two-group experiment with baseline covariate balance (left panel) and imbalance (right panel)

Consider the stylized diagrams in Figure 4. The data in the left-hand plot display good covariate balance at baseline, as evidenced by overlapping ranges of covariate values across groups and identical covariate group means (i.e., $\bar{x}_1 = \bar{x}_2 = \bar{x}$). Despite the strong positive linear relationship between the baseline covariate and the outcome, covariate adjustment will make no systematic difference to the estimates of treatment effect in cases of good balance; note that $\bar{y}_2' - \bar{y}_1' = \bar{y}_2 - \bar{y}_1$ in the left panel of Figure 4.

Data in the right-hand plot display covariate imbalance at baseline. If X and T were truly independent, we would not expect such drastic lack of overlap in covariate values across groups and differing covariate group means (i.e., $\bar{x}_1 \neq \bar{x_2}$). In this case, the covariate X appears to be a confounding variable because it is associated both with the group assignment and with the outcome variable. If the assumed linear functional forms, as specified in Equation 11, are correct, estimates based on the adjusted means will remove confounding bias due to the covariate from effect estimates; note that $\bar{y}_2' - \bar{y}_1' < \bar{y}_2 - \bar{y}_1$. The trouble is that when covariate imbalance leads to lack of covariate overlap, assumptions about correct model specification must be made based on extrapolation (i.e., in the absence of data), which is not advised.

Although lack of overlap is easy to spot when controlling for only one covariate, it is more difficult to visualize as more covariates are included. One solution is to estimate the propensity score, which serves as a unidimensional balance proxy for the full set of observed covariates (Rosenbaum & Rubin, 1983). There are a number of resources that provide guidance on best practices for the use of propensity scores for checking overlap (see, e.g., Keller & Tipton, 2016; Schafer & Kang, 2008; and Steiner & Cook, 2013).

4.4 Assumptions and Limitations

The assumptions required for valid statistical inference in ANCOVA may be summarized succinctly by the statement, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

1. *Errors are statistically independent.* That is, after accounting for the variation explained by the covariates and by group membership, residual variations are assumed to be independent. This assumption could be violated if, for example, participants were repeatedly measured. In that case, each participant's set of measurements would likely be more highly correlated with one another than with those of other participants, even after controlling for fixed effects in the model. It is typically not possible to test the independence assumption with data; instead, logical arguments based on how the data were collected must be made to defend it.

2. *Errors are normally distributed with mean 0 and constant variance.* For any combination of covariate values and/or group indicators, outcome data are assumed to have been generated from identical normal distributions, each with mean 0 and constant variance, σ^2.

   (a) *Linearity.* The assumption of zero conditional mean implies that the functional form given in Equation 11 is correctly specified. This can go wrong if the true relationship between the covariate and the outcome is nonlinear or if there is effect modification (i.e., the magnitudes and/or directions of treatment effects vary with the values of a covariate).

   (b) *Constant variance.* When group sizes are equal, or nearly so, ANCOVA-based tests are generally robust to violations of the constant variance assumption. For unbalanced group size pairings, however, heterogeneous variances may cause problematic departures from expected type I error rates. The Welch-Satterthwaite correction is recommended for such cases because it maintains nominal type I error rates in the face of group variance heterogeneity and still is quite powerful relative to alternatives (Tomarken & Serlin, 1986).

   (c) *Normal distribution.* The ANCOVA omnibus null hypothesis test is somewhat robust to non-normally distributed residuals, in particular as sample sizes grow larger. Violation of the distributional assumption may invalidate the nominal type I error rate for smaller samples, especially when coupled with unbalanced group sizes.

When verifying the tenability of the latter three assumptions with data, we recommend using a multifaceted approach that includes graphical exploration, estimation of relevant effect sizes, and the use of hypothesis tests. Other issues such as multicollinearity and influential observations can have an impact on the stability of results. For more detailed discussions on checking assumptions see, e.g., Keppel and Wickens (2004), Huitema (2011), Maxwell et al. (2018), and Fox (2008).

5 Multivariate ANOVA

Multivariate analysis of variance (MANOVA) is a generalization of ANOVA to multiple dependent variables (DVs). Paralleling Equation 3, the MANOVA model is:

$$\boldsymbol{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_j + \boldsymbol{\epsilon}_{ij},\ i = 1, \ldots, n_j, j = 1, \ldots, a, \tag{18}$$

where each term is now a p-dimensional vector, and the errors $\boldsymbol{\epsilon}_{ij}$ are assumed to be independent multivariate normal $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ variables. As in the univariate case, the treatment effects may be parameterized to satisfy $\sum_{j=1}^{a} \boldsymbol{\alpha}_j = 0$. Note that each component of the observed vector $\boldsymbol{Y}_{ij}$ satisfies the univariate model 3 and that the error components are correlated, but the overall covariance matrix $\boldsymbol{\Sigma}$ is the same for all groups.

The overall omnibus null hypothesis to be tested is that the vector of treatment effects for each outcome variable is identical across all groups:

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \cdots = \boldsymbol{\alpha}_a = 0.$$

Furthermore, the decomposition in Equation 5 still holds in its multivariate version:

$$(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}) = (\bar{\boldsymbol{y}}_j - \bar{\boldsymbol{y}}) + (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}). \tag{19}$$

Since each term in Equation 19 is a vector, it can't simply be squared; instead outer products are used, and the resulting equation consists of *matrices* of sums of squares and cross products (SSCP):

$$\begin{aligned} \text{SSCP}_{\text{T; total (corrected)}} &= \text{SSCP}_{\text{B; between}} + \text{SSCP}_{\text{W; within}} \\ \sum_{j=1}^{a}\sum_{i=1}^{n_j} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})' &= \sum_{j=1}^{a} n_j (\bar{\boldsymbol{y}}_j - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_j - \bar{\boldsymbol{y}})' + \sum_{j=1}^{a}\sum_{i=1}^{n_j} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)' \end{aligned} \tag{20}$$

The following matrices are multivariate analogues to SS_B and SS_W:

$$\boldsymbol{B} = \sum_{j=1}^{a} n_j (\bar{\boldsymbol{y}}_j - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_j - \bar{\boldsymbol{y}})' \tag{21}$$

$$\boldsymbol{W} = \sum_{j=1}^{a}\sum_{i=1}^{n_j} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)' \tag{22}$$

In the univariate case the sum of squares SS_B and SS_W are independently distributed with χ^2 distributions, and in the multivariate case $\boldsymbol{B}$ and $\boldsymbol{W}$ have independent Wishart distributions.

To test the overall omnibus null hypothesis, a measure of how *large* the matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ are is needed. There is *no unique* way to extend the F-test from Equation 7 to the multivariate case. Different researchers have proposed different multivariate analogies to the F-statistic. One of the most commonly used is Wilks' (1932) Lambda statistic:

$$\boldsymbol{\Lambda} = \frac{\det(\boldsymbol{W})}{\det(\boldsymbol{B} + \boldsymbol{W})} \tag{23}$$

As Wilks' $\boldsymbol{\Lambda}$ is a ratio of the determinants of two Wishart distributed matrices, exact results for its distribution are not available in general, but there exist relationships to the F distribution for some special cases of p and a, or an asymptotic χ^2 approximation due to Bartlett (1938). Essentially, Wilks' $\boldsymbol{\Lambda}$ arises from the likelihood ratio principle and so does its asymptotic behavior. Note that more separation among the groups will drive the $\boldsymbol{W}$ matrix closer to 0; therefore, H_0 is rejected for *small* values of $\boldsymbol{\Lambda}$. Wilks' $\boldsymbol{\Lambda}$ can also be

written using the eigenvalues $\lambda_1, \ldots, \lambda_r$ of the matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$, where $r = rank(\boldsymbol{B})$ as:

$$\boldsymbol{\Lambda} = \prod_{i=1}^{r} \frac{1}{1+\lambda_i} \tag{24}$$

As groups become more separated, $\boldsymbol{W}$ becomes closer to being a singular matrix, so the eigenvalues grow larger. Other approaches for testing H_0 due to Pillai, Lawley-Hotelling and Roy are all based on the same eigenvalues but use them in a different way. For example, Pillai's trace statistic is defined as

$$V = tr((\boldsymbol{W}+\boldsymbol{B})^{-1}\boldsymbol{B}) = \sum_{i=1}^{r} \frac{\lambda_i}{1+\lambda_i}, \tag{25}$$

and H_0 is rejected for *large* values of V. For very large samples, all four criteria are indistinguishable. For small to moderate sample sizes, some simulation-based evidence suggests that Pillai's trace is more robust to non-normality (Johnson & Wichern, 2007, p. 336).

If the multivariate omnibus null hypothesis is rejected, it is typically of interest to know which variables contributed to the rejection of the null hypothesis and which did not. Rencher and Christensen (2012, p. 195) argue that follow-up univariate ANOVA F-tests do not change the overall nominal significance level, provided that they are performed *only after rejection* of the MANOVA null hypothesis. However, this does not mean that these univariate F-tests might not suffer from inflated Type II error rates (i.e., lack of power).

This is illustrated in Figure 5, where the true population distribution is a bivariate normal with mean $(0, 0)$, variances equal to 2 and 1, and a covariance equal to 1, which results in a high correlation of $\sqrt{2}/2 \approx 0.71$. This can be thought of as the distribution of the difference between two groups, which under the null hypothesis should have mean (0, 0). The outermost red ellipse corresponds to the 95% contour of the distribution, which demarcates the acceptance and rejection regions for the multivariate null hypothesis. The shaded rectangular area is the Cartesian product of the acceptance regions of the two univariate tests.

The dot represents a potential sample with mean $(2, -1)$, which is not an unusual data point if looking at each variable separately, as shown by the fact that the point is inside the rectangle corresponding to the Bonferroni adjusted simultaneous 95% CIs for the univariate tests. Such data would result in rejection of the *multivariate* hypothesis because under the null hypothesis it is very unlikely to observe data with a large positive Y_1 mean together with a negative Y_2 mean. However, neither of the follow-up univariate hypotheses would be rejected.

6 Multivariate ANCOVA

Multivariate analysis of covariance (MANCOVA) is a combination of the MANOVA model with a multivariate regression model with one or more predictors. The inclusion of these covariates leads to *noise reduction* in the sense that the variance associated with the covariates is removed from the error variance, thus providing a more powerful test for the

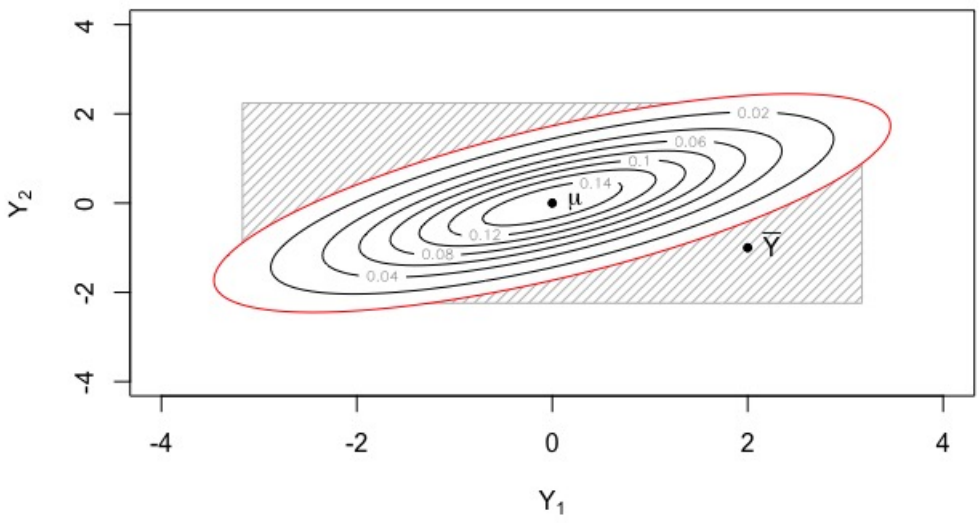

Figure 5. A bivariate normal distribution with correlation 0.71. The area outside the red ellipse represents the rejection region for the multivariate omnibus null hypothesis test. The area outside the shaded rectangle represents the Bonferroni-corrected rejection region for (either of the two) univariate omnibus null hypothesis tests. For the sample mean vector $\bar{\boldsymbol{y}}$, the multivariate test rejects while the univariate tests do not. Neither $y_1 = 2$ nor $y_2 = -1$ is an unlikely observation when *taken alone*; however, given the correlation between Y_1 and Y_2, they are quite unlikely to have been observed *together*.

difference between the groups. The model with a single covariate may be written as follows,

$$\boldsymbol{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_j + \boldsymbol{\beta}_1 x_i + \boldsymbol{\epsilon}_{ij},\ i = 1, \ldots, n_j, j = 1, \ldots, a, \tag{26}$$

where each term is a $p \times 1$ vector. The general case is easier to write in matrix notation,

$$\underset{n\times p}{\boldsymbol{Y}} = \underset{n\times a}{\boldsymbol{Z}}\ \underset{a\times p}{\boldsymbol{\alpha}} + \underset{n\times b}{\boldsymbol{X}}\ \underset{b\times p}{\boldsymbol{\beta}} + \underset{n\times p}{\boldsymbol{E}} \tag{27}$$

where $\boldsymbol{Z}$ is the MANOVA design matrix, and $\boldsymbol{X}$ is the multivariate regression matrix containing b covariates. The estimates of the coefficients are produced via standard multivariate regression formulas (see, e.g. Timm, 2002, p. 226). For example, using the generalized inverse matrix $(\boldsymbol{Z}'\boldsymbol{Z})^-$ for the case when the MANOVA is overparametrized[5], we have:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{Z}'\boldsymbol{Z})^-\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{28}$$

Observe that this estimate of the group effects is adjusted by the residual from a regression on the covariates.

[5] Following Timm (2002), we use the superscript $^-$ to denote the generalized inverse.

To test hypotheses about the treatment effects, $\boldsymbol{\alpha}$, the standard approach is to obtain the "within" matrix (now called the *residual matrix*) $\boldsymbol{W}_{full}$ from the full model in Equation 27 and the residual matrix $\boldsymbol{W}_{reduced}$ from the model under the null hypothesis, and then proceed the same way as in the MANOVA hypothesis, using the eigenvalues of $\boldsymbol{W}_{full}^{-1}(\boldsymbol{W}_{full} - \boldsymbol{W}_{reduced})$. To summarize, the MANCOVA test enjoys the benefits of both MANOVA (i.e., increased power for rejecting the omnibus test with correlated outcomes) and ANCOVA (i.e., reduction of within-groups error terms via regression).

6.1 Assumptions and Limitations

Significance tests for MANOVA, MANCOVA, and many other multivariate statistical analyses are based on the assumption that the data are multivariate normal. More specifically, the error matrix $\boldsymbol{E}$ in Equation 27 is assumed to contain on each row independent vectors which are distributed $N_p(0, \boldsymbol{\Sigma})$.

1. *Row vectors are independent.* This assumption is analogous to the univariate case with exception that the observations associated with the same participant are assumed to be correlated as specified in $\boldsymbol{\Sigma}$.

2. *Error vectors are multivariate normal with zero mean vector and constant covariance matrix.*

   (a) *Linearity.* In line with the univariate case, the relationships between covariates and outcome variables are assumed to be correctly specified.

   (b) *Constant covariance matrix.* The assumption that each error vector is distributed $N_p(0, \boldsymbol{\Sigma})$ implies identical covariance matrices. This assumption is the multivariate analogue to the univariate assumption of constant variance. The assumption may be tested with Box's M test, though the test is sensitive to violations of other assumptions such as multivariate normality.

   (c) *Multivariate normal distribution.* Multivariate normality implies that each outcome variable must be normally distributed, which can be tested with a number of tests such as Shapiro-Wilks, Anderson-Darling or Kolmogorov-Smirnov, as well as graphs such as a Q-Q plot. However, multivariate normality is a stronger assumption than individual univariate normality, and there are multivariate tests available in many software packages such as Mardia's test and the Cox-Small test. Mardia (1971) argues that reasonably large sample sizes in each group ensure robustness and mitigates the need for extensive testing, though Rencher and Christensen (2012, p. 210) suggest checking for gross violations.

MANOVA and MANCOVA procedures are sensitive to outliers. Thus, it is recommended that tests for outliers are performed with any MANOVA and MANCOVA analysis. For more discussion of assumptions and their implications see, e.g., Rencher and Christensen (2012), Johnson and Wichern (2007) and Tabachnick and Fidell (2013).

6.2 Data Demonstration

Woodworth, O'Brien-Malone, Diamond, and Schüz (2017) replicated a randomized efficacy study of three web-based positive psychology interventions and a control condition, and Woodworth, O'Brien-Malone, Diamond, and Schüz (2018) made the data publicly available. Participants assigned to the *gratitude* condition were asked to send a letter of gratitude to someone deserving in their life. Participants assigned to the *three good things* condition were asked to write down three good things that happened each day for a week. Participants assigned to the *signature strengths* condition completed a survey to assess character strengths and were asked to use one of their top five strengths each day for a week. Participants in the comparison group were asked to write about early memories each day for a week.

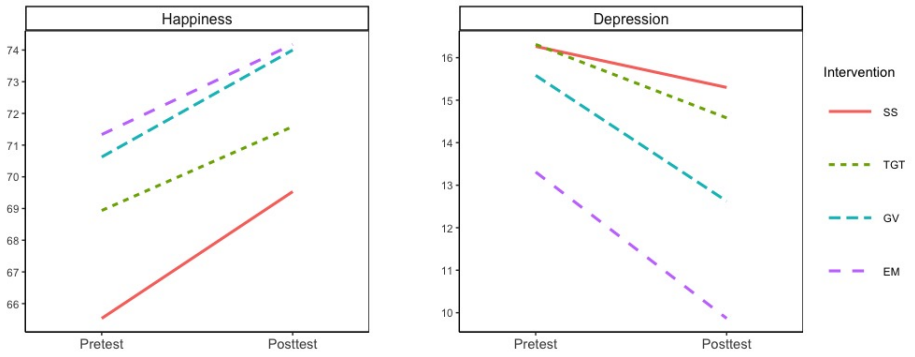

Figure 6. Pretest and posttest scores on two dependent variables for the data demonstration; note the lack of balance across intervention group means at baseline

Of interest is the effect of the intervention type ($a = 4$ groups) on the depression and happiness index scores ($p = 2$ DVs, correlated at -0.76). For this analysis we focus on the 147 participants who completed baseline and posttest measures. Neither Wilks nor Pillai MANOVA omnibus tests are significant ($p = 0.28$). A follow-up analysis with univariate F-tests on each of the DVs separately agrees with this result. Running a MANCOVA, controlling for baseline scores as covariates, also results in a nonsignificant omnibus test result ($p = 0.45$) via both Wilks and Pillai. The error matrices for the MANOVA and MANCOVA are:

$$\boldsymbol{W}_{MANOVA} = \begin{pmatrix} 17736.79 & -15240.07 \\ -15240.07 & 22637.71 \end{pmatrix}, \; \boldsymbol{W}_{MANCOVA} = \begin{pmatrix} 8458.66 & -5233.99 \\ -5233.99 & 7714.14 \end{pmatrix}$$

As expected, adjusting for baseline scores reduced important components of the error variance matrix. It is instructive to note that $det(\boldsymbol{W}_{MANCOVA}) = 3.8 \times 10^7$ is 4.5 times smaller than $det(\boldsymbol{W}_{MANOVA}) = 1.7 \times 10^8$. A similar comparison of the "Between" matrices (for formulas see Timm, 2002, p. 227), yields

$$\boldsymbol{B}_{MANOVA} = \begin{pmatrix} 724.04 & -525.16 \\ -525.16 & 481.89 \end{pmatrix}, \; \boldsymbol{B}_{MANCOVA} = \begin{pmatrix} 216.39 & -23.68 \\ -23.68 & 18.7 \end{pmatrix}$$

In general, for randomized experiments, we do not expect to see large differences in

the MANOVA and MANCOVA covariate-adjusted between matrices. In this example, however, the determinant of the between matrix decreased by a factor of almost 21. The explanation for this large reduction in the magnitudes of the components of the between matrix is the lack of baseline balance in the analytic sample, perhaps due to post-randomization attrition. MANOVA fails to account for the baseline differences and attributes all outcome variation to the treatment. Whereas, with MANCOVA, the baseline variation in outcomes across treatments is accounted for and, thus, $det(\boldsymbol{B})$ is substantially smaller. Note that we use these data to demonstrate the mechanics of the multivariate procedures. We have skipped over important details such as the verification of the tenability of assumptions and the reporting of and interpretation of results in context that should not be neglected in practice.

7 Discussion

Random assignment leads to covariate balance, which renders both ANOVA and ANCOVA-based estimators of effects unbiased (see the left panel of Figure 4). The purpose of making ANCOVA adjustments with randomized experimental data is to reduce error variance, which leads to more precise estimates and more powerful tests. With observational data, ANCOVA may still be useful in reducing error variance, but it also is used for another purpose: to reduce bias due to confounding covariates.

The ANCOVA and MANCOVA models, as described in Equations 11 and 26, require (a) that relationships between covariates and outcomes are linear and (b) that treatment effects are constant across all covariate values. If either of these assumptions is violated, estimates based on adjusted means are no longer valid. Furthermore, because extrapolation is required when covariate imbalance leads to lack of overlap, the validity of the estimates may be quite sensitive to functional form misspecifications with observational data.

That said, some empirical evidence suggests that in some cases the basic ANCOVA model can perform quite well with respect to bias reduction with observational data (Shadish, Clark, & Steiner, 2008); however, simulation evidence suggests that when functional form assumptions are violated, performance may be substantially improved by allowing for more flexible modeling, for example, by including covariate by treatment interaction terms to handle non-constant treatment effects (Schafer & Kang, 2008).

There is a rich and expanding literature on alternatives to ANCOVA for effect estimation in observational study settings in the causal inference literature; some of these approaches are reviewed in Schafer and Kang (2008). In recent years, a number of machine learning alternatives to ANCOVA for causal effect estimation with observational data have been proposed. These newer methods lack the simplicity and interpretability of ANCOVA but benefit from algorithmic handling of nonlinearities and interactions in their naive implementations. Bayesian additive regression trees (Hill, Weiss, & Zhai, 2011) and targeted maximum likelihood estimation (TMLE; van der Laan & Gruber, 2010) are two examples that have performed well in data competitions (Dorie, Hill, Shalit, Scott, & Cervone, 2019). Furthermore, a number of new methods have been proposed for explicitly detecting and estimating causal effects in the presence of treatment effect heterogeneity. A special issue in the journal *Observational Studies* (Volume 5, Issue 2, 2019) is devoted to the topic.

Our focus here has been on one-way (i.e., single-factor) models for ANCOVA and

MANCOVA. Of course, it may be of interest to adjust for covariates in experimental contexts involving two or more factors and their interactions. It poses no problem to extend to multifactor studies, and a number of texts provide relevant test statistics and formulas for adjusted cell means in the univariate case (e.g., Neter et al., 1990; Kirk, 1982). In the multivariate case, one would need to alter the MANOVA design matrix, i.e., $\boldsymbol{Z}$ in Equation 26, to include the appropriate deviation-coded variables; see Timm (2002, p. 246) for details.

Finally, after a rejection of the MANCOVA omnibus null hypothesis, as in the multivariate data demonstration given above, univariate ANCOVA F tests are not the only options for following up to better understand what contributed to the rejection. Instead, for example, one might choose to search for the linear combination of outcome variables that lead to the best separation among treatment groups. This approach, called *discriminant analysis*, is beyond the scope of this chapter but is covered in the multivariate texts referenced herein.

8 References

Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Mathematical Proceedings of the Cambridge Philosophical Society*, *34*, 33–40.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*, 43–68.

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). London, England: Oliver and Boyd.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high–dimensional setting and a potential alternative. *Multivariate Behavioral Research*, *46*, 477–513.

Huitema, B. (2011). *Analysis of covariance and alternatives* (2nd ed.). New York: Wiley.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics*, *41*, 326-348.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, *58*, 105–121.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). New York: Routledge.

Morgan, K. L., & Rubin, D. B. (2014). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*, 1263–1282.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models: Regression, analysis of variance, and experimental designs* (3rd ed.). Burr Ridge, IL: Irwin.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Schafer, J., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Shadish, W. R., Clark, M. H., & Steiner, P. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1334–1343.

Steiner, P. M., & Cook, T. D. (2013). Matching and propensity scores. In T. Little (Ed.), *Oxford handbook of quantitative methods.* Oxford: Oxford University Press.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90–99.

van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*, Article 17.

Wildt, A. R., & Ahtola, O. T. (1978). *Analysis of covariance.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 12. Beverly Hills and London: Sage Publications.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, *24*, 471–491.

Woodworth, R. J., O'Brien-Malone, A., Diamond, M. R., & Schüz. (2017). Web-based positive psychology interventions: A reexamination of effectiveness. *Journal of Clinical Psychology*, *73*, 218–232.

Woodworth, R. J., O'Brien-Malone, A., Diamond, M. R., & Schüz. (2018). Data from, 'web-based positive psychology interventions: A reexamination of effectiveness'. *Journal of Open Psychology Data*, *6*, 1.

Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B., & Zhang, A. (2018). Sequential rerandomization. *Biometrika*, *105*, 745–752.