A Comparison of Parametric and Permutation Tests for Regression Analysis of Randomized Experiments

Kellie Ottoboni
Department of Statistics; Berkeley Institute for Data Science
University of California, Berkeley

Fraser Lewis
Medical Affairs and Evidence Generation
Reckitt Benckiser

Luigi Salmaso
Department of Management and Engineering
University of Padova

August 24, 2017

Abstract

Hypothesis tests based on linear models are widely accepted by organizations that regulate clinical trials. These tests are derived using strong assumptions about the data-generating process so that the resulting inference can be based on parametric distributions. Because these methods are well understood and robust, they are sometimes applied to data that depart from assumptions, such as ordinal integer scores. Permutation tests are a nonparametric alternative that require minimal assumptions which are often guaranteed by the randomization that was conducted. We compare analysis of covariance (ANCOVA), a special case of linear regression that incorporates stratification, to several permutation tests based on linear models that control for pretreatment covariates. In simulations using a variety of data-generating processes, some of which violate the parametric assumptions, the permutation tests maintain power comparable to ANCOVA. We illustrate the use of these permutation tests alongside ANCOVA with data from a clinical trial comparing the effectiveness of two treatments for gastroesophageal reflux disease. Given the considerable costs and scientific importance of clinical trials, one may want to include an additional nonparametric method, such as a linear model permutation test, as a robustness check on the statistical inference for the main study endpoints.

Keywords: Nonparametric methods, linear model, analysis of covariance, analysis of designed experiments

1 Background

A hypothesis test is a statistical method for determining whether observed data is consistent with a belief about the process that generated the data. Medical experiments use hypothesis testing to assess the evidence that a treatment affects one or more clinically relevant outcomes. The simplest version of this experiment, which involves randomly assigning two treatments to a fixed number of individuals in a group and measuring a single outcome, has been studied for nearly a century (see Fisher (1935) and (Neyman, 1923, 1990 translation) for early references). One can conduct hypothesis tests and construct confidence intervals for the estimated treatment effect by exploiting the fact that the difference in average outcomes between the two treatment groups is asymptotically normal with a variance that can be estimated from the data.

Randomized experiments in the real world are rarely this simple: if individuals are heterogeneous before the study, then their outcomes may differ for reasons besides the treatment; there may be more than two treatments under study; patients may not take the treatment they are assigned or may drop out of the study; and peer effects, the effect of being in the same treatment group as others, may affect outcomes on top of the treatment effect alone. Techniques have been developed to deal with each of these issues, invoking additional assumptions in order to carry out valid inference. In this paper, we focus on the issue of heterogeneity of pretreatment characteristics and assume that the other issues are not present.

Random assignment of treatment ensures that pretreatment covariates are balanced between treatment groups on average, across all possible randomizations. However, in any particular randomization, there may be imbalances. If the imbalanced variables are associated with the outcome, then even when treatment has no effect, there may be differences in outcomes between treatment groups. Adjusting for such covariates can reduce the variability of treatment effect estimates and yield more powerful hypothesis tests.

Stratification is one method to control for covariates that are known a priori to be associated with the outcome. Strata are groups of individuals with similar levels of a covariate. These groups are defined during the design stage (i.e. before outcome data are collected). Random assignment of treatments is conducted within each stratum, indepen-

dently across strata. This guarantees that the stratification variable is balanced between treatment groups. A common stratification variable in clinical experiments is location: individuals often come from many sites because it is difficult to recruit a sufficient number of participants at one doctor's office or hospital, especially when the object of study is a rare disease or a rare outcome. Controlling for stratum can have the same beneficial effects as controlling for other pretreatment covariates, as outcomes may vary across strata due to factors unrelated to individuals' treatment and covariates.

Linear regression is another way to control for baseline covariates. It is done during the data analysis stage. While stratification guarantees balance on the stratifying variable, other covariates may still be imbalanced. Stronger modeling assumptions may be necessary to balance them after the experiment. Linear regression is a method which projects the outcomes onto the plane that best summarizes each variable's relationship to the outcome. The coefficient of any particular covariate answers the question, if we were to hold fixed all other variables and increase this variable by one unit, how much would we expect the outcome to change? This model posits a linear relationship between covariates, treatment, and outcome; if the true relationship is not linear, then linear regression gives the best linear approximation to the conditional expectation of the outcome. When controlling for strata, such as location, it is standard to use analysis of covariance (ANCOVA). ANCOVA is a particular case of linear regression that allows the mean outcome to vary from stratum to stratum. This amounts to fitting a plane to each stratum, with the constraint that they have a common slope.

Hypothesis testing of estimated coefficients requires even stronger assumptions than linearity, namely Gaussian, homoskedastic errors. When this holds, a regression coefficient scaled by its estimated variance follows Student's t distribution. Thus, if we believe the linear model is correctly specified, a hypothesis test for a treatment effect amounts to a hypothesis test of the coefficient for treatment in the ANCOVA model, and can be evaluated analytically and efficiently. However, the modeling assumptions are clearly violated in many cases: when the data are discrete or ordinal, when the treatment has a differential effect across subgroups of individuals and no interaction terms are included, and when the variance of outcomes differs across strata. Furthermore, the hypothesis test implicitly

assumes that the data were sampled at random from some underlying population, when in fact, medical experimenters rarely recruit patients this way (Ludbrook and Dudley (1998)).

Permutation testing is an alternate approach (Fisher (1935); Pitman (1937, 1938)). Deliberate randomization induces a distribution for any test statistic under the null hypothesis that treatment has no effect on the outcome: the randomization scheme provides information about all possible ways that treatment may have been assigned and the null hypothesis tells us what each individual's response would be regardless of the assignment (namely, it would be the same). One determines how "extreme" the observed test statistic is relative to this randomization distribution, rather than a parametric reference distribution like Student's t or the standard Gaussian. Such a test is exact, meaning that it controls the type I error rate at the pre-specified level even in finite samples, whereas parametric hypothesis tests based on asymptotic approximations do not always guarantee good finite sample properties. Permutation tests condition on the observed sample and do not require any assumptions about the way individuals were sampled from a larger population. This is useful when the sampling frame is difficult to specify, such as when individuals are a convenience sample. While external validity may be an issue, we may still be interested in whether the treatment had an effect on the observed individuals. TO DO: CLEAR?

In the past, statisticians relied on parametric methods because asymptotic approximations were a computationally feasible way to estimate distributions and construct confidence intervals. Now, computational power is no longer a barrier to finding exact (or exact to pre-specified precision) randomization distributions and confidence intervals. In most cases, a randomization test is the "gold standard": "[a] corresponding parametric test is valid only to the extent that it results in the same statistical decision [as the randomization test]" (Bradley (1968)). If the permutation test agrees with the parametric test, one may have a greater degree of confidence in the estimates and confidence intervals constructed using the parametric method.

There is no hard and fast rule describing the rate at which parametric tests approach the exact permutation solution, as they are both highly dependent on the particular data observed. On the one hand, some argue that violations of parametric test assumptions necessitate the use of permutation methods. Ludbrook and Dudley (1998) point out that

medical trials rarely follow the population sampling model implicit in parametric methods. Many people recommend using permutation tests in place of the usual parametric tests for typical analyses, such as ANOVA and generalized linear models (Still and White (1981); Winkler et al. (2014)). They argue that there are a myriad of ways that the data may violate the necessary assumptions for the test, and so permutation tests are more robust.

However, parametric and nonparametric tests seem to perform similarly when compared side-by-side in simulations, even when data violate the assumptions of the parametric method. Medical trials often use Likert scales to score pain or symptom severity, resulting in discrete data that does not match the normality assumptions of parametric tests. However, de Winter and Dodou (2010) found that the two sample t test and Mann-Whitney test had comparable Type I and II error rates for five-point Likert scale data, suggesting that the violation of normality does not entirely invalidate the parametric test. Vickers (2005) compared the parametric ANCOVA to the Mann-Whitney rank test in the context of randomized experiments, finding that except in extreme situations, ANCOVA was more powerful than the nonparametric test. Most similar to our question of study, Anderson and Legendre (1999) found little difference between several permutation tests for coefficients in a linear model alongside the parametric t test. In these situations, the permutation test strengthens conclusions by giving evidence that the parametric test is robust to departures from its assumptions.

In this paper, we review several hypothesis tests for a treatment effect which adjust for pretreatment covariates to increase power. We focus on ANCOVA and its permutation counterparts, comparing their performance in different scenarios and illustrating their application with a clinical dataset. Section 2 describes the various tests mathematically and states their assumptions explicitly. Section 3 presents simulations that suggest that even when assumptions are not satisfied, the parametric and permutation tests have comparable power to detect a treatment effect. In Section 4, we apply each of the tests to data from a clinical trial comparing the performance of two treatments for gastroesophageal reflux disease (GERD). We conclude in Section 5 with implications of these results for practitioners.

2 Methods

2.1 Notation

Suppose we randomly assign two treatments labelled 0 and 1 to a group of individuals. We are interested in comparing the relative effectiveness of treatment 1 to treatment 0. Let Z indicate treatment assignment, Y denote the observed outcome of interest, and X be a pretreatment covariate that is observed and associated with the outcome. For expository clarity, we suppose that X is univariate, but all results are easily extended to the case when X is multivariate. Suppose further that there is a categorical pretreatment variable with J levels used to stratify individuals.

Strata are indexed by subscript $j = 1, \ldots, J$. Let n_j be the number of individuals in stratum j. Individuals within stratum j are indexed by subscript $i = 1, \ldots, n_j$. We observe $\{Y_{ij}, Z_{ij}, X_{ij}\}$ for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$. All individuals have two potential outcomes, $Y_{ij}(1)$ and $Y_{ij}(0)$, representing their responses to treatments 1 and 0, respectively. We can never observe both; random assignment of treatment reveals $Y_{ij} = Z_{ij}Y_{ij}(1)+(1-Z_{ij})Y_{ij}(0)$. Throughout, we assume that there is no interference between individuals (in other words, Y_{ij} is a function of $(Z_{ij}, Y_{ij}(1), Y_{ij}(0))$ and not any other $Z_{i'j'}$ for $(i', j') \neq (i, j)$.) and that there is no censoring or non-compliance (we actually observe $Y_{ij} = Y_{ij}(Z_{ij})$ for all (i, j)).

We are interested in the differences in potential outcomes $Y_{ij}(1) - Y_{ij}(0)$. This quantity represents the effect of treatment for individual i in group j: it is the difference between what we would have observed under treatment 1 and what we would have observed under treatment 0. We may never learn this difference for any particular individual, so a typical problem is to estimate the mean difference in some group, such as the study sample or in a target population. The chosen method of analysis determines which function of differences is considered. Various functions of potential outcomes may be of clinical interest; which one prefers depends on the goal of the study.

We study hypothesis testing for whether these differences are nonzero using parametric ANCOVA and its permutation counterparts, assuming this potential outcomes framework throughout. TO DO: MAKE SURE THE PARAMETRIC SECTION ACTUALLY REFLECTS THE POTENTIAL OUTCOMES FRAMEWORK. SEE JAS'S EMAIL AND THE WINSTON LIM PAPER

Other valid methods for comparing two groups include using a two-sample t-test to test the difference in two means from normal distributions, the Wilcoxon rank sum test to test for differences in the medians of two independent groups, and the Kolmogorov-Smirnov test and receiver operating characteristic curve analyses to test whether two groups have different distribution functions (Lehmann (1975); Vexler et al. (2016)). We focus on testing using the linear model as this is standard in clinical trials, requires fewer distributional assumptions on the data when using the potential outcomes framework, and incorporates control variables to increase power.

2.2 Parametric ANCOVA

ANCOVA is based on a linear model with an indicator variable for membership in each stratum. The model is

$$Y_{ij} = \alpha_j + \beta X_{ij} + \gamma Z_{ij} + \varepsilon_{ij} \tag{1}$$

where α_j is a fixed effect for stratum j, β is the coefficient for the pretreatment covariate, γ is the coefficient for treatment, and ε_{ij} is an error term. The parameter of interest is γ. If we believe the linear model is the true data-generating process, it asserts that for each individual, $Y(1) = Y(0) + \gamma$. However, we needn't take this perspective for γ to be a useful quantity; it represents the average treatment effect, holding the other variables fixed. We would like to test the null hypothesis $H_0 : \gamma = 0$ against the two-sided alternative hypothesis $H_1 : \gamma \neq 0$.

To carry out the standard parametric hypothesis test for a linear model, we make the following assumptions (Freedman (2005)):

1. **Linearity:** The data Y are related to X and Z linearly.

2. **Constant slopes:** Stratum membership only affects the intercept α_j, not the slopes β and γ.

3. **IID Errors:** The ε_{ij} are independent and identically distributed with mean 0 and common variance σ^2.

4. **Independence:** If X is random, ε is statistically independent of X.

5. **Normality:** The errors are normally distributed.

TO DO: ARE THE ESTIMATOR, STANDARD ERROR, VARIANCE CLEAR HERE? The coefficients are estimated using least squares. The estimated coefficient $\hat{\gamma}$ is the estimated average treatment effect. (Note, only Assumptions (1) and (2) are needed to guarantee the existence of a solution. The solution will be unique as long as there is no linear relationship between X, Z, and stratum membership.) This procedure also yields an estimate $\hat{\sigma}^2_{\hat{\gamma}}$ of the variance of $\hat{\gamma}$ (namely, the corresponding diagonal element of the inverse covariance matrix of regression covariates scaled by the residual sum of squares over the degrees of freedom). Under the null hypothesis, the test statistic

$$T = \frac{\hat{\gamma}}{\sqrt{\hat{\sigma}^2_{\hat{\gamma}}}}$$

follows the Student t distribution with degrees of freedom equal to the number of observations minus the number of parameters estimated (in this case, $N - J - 2$). The p-value for this hypothesis test is the probability, assuming the null hypothesis of zero coefficient is true, that a value drawn from the t distribution is larger in magnitude than T. This test is equivalent to the F-test and the likelihood ratio test assuming the Gaussian model, and TO DO: CITE THIS? IS THEREFORE LOCALLY MOST POWERFUL ????.

The ANCOVA model includes a fixed effect for each stratum, but the model does not account for variation in the treatment effect across strata. The estimated linear model may not accurately capture details of the true data-generating process: if effects are not constant across strata, then the coefficient γ may be attenuated towards 0. A test of the null hypothesis that $\gamma = 0$ will be valid, but will not reflect the true magnitude of treatment effects among individuals. One can account for differential treatment effects across strata by including interaction terms for treatment and stratum membership, but this complicates the hypothesis testing problem.

2.3 Stratified permutation test

Suppose we wish to test the null hypothesis that individual by individual, treatment has no effect. This is referred to as the "sharp" null hypothesis:

$$H_0 : Y_{ij}(1) = Y_{ij}(0), \forall i = 1, \ldots, n_j, j = 1, \ldots, J.$$

Then, which treatment an individual received amounts to an arbitrary label. Once we observe one response under a particular treatment, we can impute the other potential outcome; namely, it would have been the same. This null hypothesis is stronger than the null hypothesis for the parametric ANCOVA, which only asserts that the treatment has no effect *on average*.

For the permutation test, we condition on the number of individuals who received each treatment within each stratum. Any assignment of treatments that preserves the number of treated units within each stratum is valid and was just as likely to occur as the randomization that was actually observed. We can construct the permutation distribution of any statistic under the null hypothesis using this principle of equal probabilities and by imputing the unobserved potential outcomes using the sharp null hypothesis.

The most commonly used statistic is the difference in mean outcomes of subjects who received treatment 1 and the mean outcomes of subjects who received treatment 0. This statistic is unbiased over all possible random assignments of treatment which preserve the number of treated individuals, is interpretable ("on average, taking treatment changes the outcome by x amount"), and has convenient theoretical properties owing to it being the difference of two means. However, the difference in means may not be optimal if we want to be sensitive to heterogeneous effects. For an extreme example, imagine that the sample contains an equal number of males and females, and each treatment is assigned to half of males and half of females. Everybody who receives treatment 0 has an outcome of 0, but males who receive treatment 1 have an outcome of 1 and females who receive treatment 1 have an outcome of -1. Then the difference in means between the treatment groups is 0, even though the treatment had nonzero effects on both males and females. This differential effect gets averaged out.

To avoid this, one may want to stratify the sample according to important confounding variables, then compute a statistic for each stratum separately. There is a great degree of freedom in deciding how to construct such a statistic. Two choices must be made: how to stratify and how to combine the statistics across strata. There is a tradeoff in how

finely we choose to stratify. On the one hand, each stratum must be sufficiently large to be informative, but on the other hand, must be fine enough to capture variation in treatment effects. Our simulations in Section 3 show that power decreases when effects are concentrated in small strata.

One way to construct a test statistic is to directly combine the stratum statistics into a single value, for instance taking the sum of their absolute values. Taking the absolute value before summing ensures that effects with different signs do not cancel each other out. Permutations for this test are conducted identically to those in the previous section, but this combined statistic is used in place of the difference in means. Another way to construct a test is to use the nonparametric combination (NPC) framework proposed by Pesarin and Salmaso (2010). In this framework, one would break down the "global" null hypothesis of no effect whatsoever into the intersection of "partial" null hypotheses of no treatment effect within each stratum. For NPC, one first conducts each partial test separately, then combines their p-values in a way that preserves dependencies. In randomized experiments, strata are randomized independently, making this method equivalent to combining stratum statistics directly.

2.4 Permutation tests with the linear model

The tests in Section 2.3 only use information on the treatment and the outcome. However, experimenters record additional covariates, some of which may be predictive of the outcome. One would like to construct a more powerful test by incorporating these covariates to reduce the variance of the statistic. The permutation tests in this section continue to model the outcome as in Equation 1, but require fewer assumptions about the data. In a randomized experiment, the treatment assignment is independent of covariates, errors, and potential outcomes, making several variables "permutable." We show two permutation tests developed in this framework.

First, we may do a simple variation on the stratified permutation test described above. Instead of using the difference in means as the statistic, we use the t statistic for the coefficient on Z in the linear regression. If treatment was assigned at random within each stratum, then it is guaranteed that Z and ε are statistically independent, conditional on

stratum. Therefore, we can construct a permutation distribution by repeatedly permuting treatment assignments Z within strata, independently across strata, and regressing the outcomes on the covariate and permuted treatment vectors.

Freedman and Lane (1983) propose an alternative test based on the residuals of the linear regression. They take an alternative view of the problem, still writing the outcome in the form of Equation 1, but they *define* the errors ε_{ij} to be the difference between the observed outcome Y_{ij} and the data's linear projection onto the plane $\alpha_j + \beta X_{ij} + \gamma Z_{ij}$. The ε are fixed approximation errors in this framework, not independent and identically distributed random errors.

If the null hypothesis is true, then $\gamma = 0$ and $\varepsilon_{ij} = Y_{ij} - \alpha_{ij} - \beta X_{ij}$ for all $i = 1, \ldots, n_j, j = 1, \ldots, J$. Therefore, we may estimate the errors $\hat{\varepsilon}$ by $Y - \hat{Y}$, where $\hat{Y}$ is the vector of fitted values from the regression of Y on X but not Z. The $\hat{\varepsilon}$ approximate the true errors ε from the true data-generating process, assuming that the null hypothesis is true and the linear model has a reasonable in-sample fit. (Freedman and Lane (1983) advise against using this method if there are large outlier values in the covariate X or when X and Z are highly colinear.) TO DO: NOT CLEAR HOW TO FORMALLY TEST THE CONDITIONS FOR THE PERMUTATION TEST. VISUAL INSPECTION IS NOT SATISFACTORY Furthermore, under the null hypothesis, the ε are independent Z within strata. If we make the additional assumption that ε are exchangeable against X within strata, then we may permute the estimated $\hat{\varepsilon}$ within strata, independently across strata. This additional exchangeability assumption should be checked using the residuals $\hat{\varepsilon}$; for instance, visual inspection of a scatterplot of the residuals against X should not reveal any patterns if they are exchangeable.

To summarize the steps for constructing a permutation distribution:

1. Estimate $\hat{\varepsilon}$ by $Y_{ij} - \hat{\alpha}_j - \hat{\beta} X_{ij}$ for all i and j, where $\hat{\alpha}_j$ and $\hat{\beta}$ are obtained by regressing Y on X and stratum assignment, *but not* Z.

2. Obtain permuted errors $\hat{\varepsilon}^{\pi}$ by permuting the $\hat{\varepsilon}$ within strata.

3. Compute permuted responses $Y^{\pi}_{ij} = \hat{\alpha}_j + \hat{\beta} X_{ij} + \hat{\varepsilon}^{\pi}_{ij}$.

4. Regress Y^π on X, Z, and stratum. The test statistic is the t statistic for the coefficient of Z.

Others have developed variants on these approximate regression-based permutation tests. There is some disagreement on what constitutes an appropriate permutation scheme; everything we have described so far is approximate insofar as we do not know the "true" data-generating process described by the α_j and β, only the estimates $\hat{\alpha}_j$ and $\hat{\beta}$. Manly (2006) proposed randomizing the outcomes Y, treating them as though they were randomly assigned to pairs (X_{ij}, Z_{ij}) under the null hypothesis. TO DO: IS THIS CLEAR? Permutations under the null hypothesis assume that all coefficients are 0 in the true data-generating process, which may not reflect the true relationship between variables. Kennedy (1995) proposed a permutation method similar in spirit to Freedman and Lane (1983), but which differs procedurally. Both methods attempt to measure the correlation between treatment and unexplained variation in outcomes, but instead of regressing pseudo-outcomes Y^π on covariates to obtain a permutation distribution, Kennedy (1995) recommends using the t statistic from regressing Z on the permuted residuals $\hat{\varepsilon}^\pi$.

Several authors have compared these tests theoretically and empirically (Anderson and Legendre (1999); Anderson and Robinson (2001); Kennedy and Cade (1996)). They find that the Freedman-Lane test of residuals is asymptotically equivalent to the "oracle" exact hypothesis test which we could conduct if we knew which permutations of Y given X were equiprobable under the null hypothesis (Anderson and Robinson (2001)). This confirms empirical results, which show that the Freedman-Lane test performs better than other linear regression based tests in simulations (Anderson and Legendre (1999)), though its advantage is small. Therefore, throughout the rest of the paper we focus on the two linear regression permutation tests we described in detail: linear regression with permuted treatment assignments and the Freedman-Lane method of permuting residuals.

3 Simulations

We simulated data from a randomized experiment using several different data-generating processes. We applied the tests to the data and compared their empirical power over re-

peated random treatment assignments and random errors. We compared the following five tests: the t test from the parametric ANCOVA, a stratified permutation test using the difference in mean outcomes (called "stratified permutation" in what follows), a stratified permutation test using the mean change scores, defined as the difference between the baseline measure and the outcome (called "differenced permutation" in what follows), a stratified permutation test based on the t statistic from the linear regression of outcome on control variables (called "LM permutation" in what follows), and the Freedman-Lane permutation test. The simulations are included in three supplementary files.

Outcomes were generated according to the linear equation

$$Y_{ij1} = \alpha_j + \beta_0 Y_{ij0} + \gamma_j Z_{ij} + \varepsilon_{ij} \tag{2}$$

for individuals $i = 1, \ldots, n_j$, $j = 1, 2, 3$. α_j is the mean effect of being in stratum j, β_0 is the coefficient for the baseline measurement Y_{ij0}, Z_{ij} is the treatment level, γ_j is the effect of treatment in stratum j, and ε_{ij} is an error term. We used three strata with $\alpha_1 = 1, \alpha_2 = 1.5$, and $\alpha_3 = 2$.

We used two designs:

- Constant additive treatment effect: $\gamma_1 = \gamma_2 = \gamma_3 = \gamma$. This is the implicit assumption when using a linear model.

- Treatment effect in a single stratum: $\gamma_1 = \gamma > 0$, $\gamma_2 = \gamma_3 = 0$. This is a constant, additive treatment effect in stratum 1, but no treatment effect in strata 2 and 3. This is a simplistic example of a heterogeneous treatment effect. The standard ANCOVA model does not account for this scenario.

We drew the baseline measurements once, treating them as fixed, and conditioned on observing $\{Y_{ij0} : i = 1, \ldots, n_j, j = 1, \ldots, 3\}$. We randomly assigned treatment to half of the individuals within each stratum, generated random errors, and constructed new outcomes Y_{ij1} using Equation 2. We conducted all five tests on this new data, obtaining a two-sided p-value for each. We repeated these steps using the same $\{Y_{ij0}\}$ 1000 times.

In our first set of simulations, we assumed that the baseline measurements Y_{ij0} were standard normally distributed and the pairs $(Y_{ij0}, \varepsilon_{ij})$ were independent across i and j. We

let $\beta_0 = 1$ and the treatment effect be $\gamma = 1$. We used a balanced design with $n_j = 16$ individuals per stratum and balanced treatment assignment, i.e. 8 subjects received each treatment within each stratum. There were four total simulation designs: for each of the two possible treatment effects, we used two distributions for ε. In the first case, we used $\varepsilon \sim N(0, 1)$ to match the parametric assumption of normal errors. In the second case, ε were drawn from a t distribution with 2 degrees of freedom so the errors were heavy-tailed.

Figure 1 shows the estimated power curves in these four designs. The best case was when the errors were Gaussian and the treatment effect was constant across strata, while the worst case was when the errors are heavy-tailed and the treatment effect only appeared in one stratum. Intuitively, it makes sense that power decreased relative to the Gaussian, constant treatment effect case, as each violated assumption further obscured the treatment effect. A consistent pattern appeared in each design: the stratified permutation test had the lowest power, while the other four tests tended to have similar power.

Table 1 displays the size of the test and the power at level 0.05. The three tests based on linear models had slightly higher than nominal level, though the margin was within two standard errors of 0.05. (The number of tests rejected under the null in 1000 trials has a binomial distribution. If the true level is 0.05, then the standard error is $\sqrt{0.05 \times 0.95/1000}$.) These numbers show that actually, for $\alpha = 0.05$, the stratified permutation test did not lose a large amount of power when the effect was isolated in a single stratum. In this case, the stratified permutation test may have slightly higher power than ANCOVA at the small significance levels typically used in practice.

Our second set of simulations used the same design as the previous ones, but modified the sizes of the treatment groups. Within each stratum, 4 individuals received one treatment and 12 received the other. This mimics some real-world experiments, where the more expensive treatment is administered less frequently than the placebo or standard of care. It is well-known that both parametric and nonparametric tests have higher than nominal level when there is heterogeneous variance, and this effect is exacerbated when group sizes are unequal (Glass et al. (1972), Zimmerman (2006)). Here, the data had equal variances so we would expect power, not the test level, to be an issue. (See the supplementary Gaussian simulations for simulations using data with heterogeneous variances.) Table 2 displays

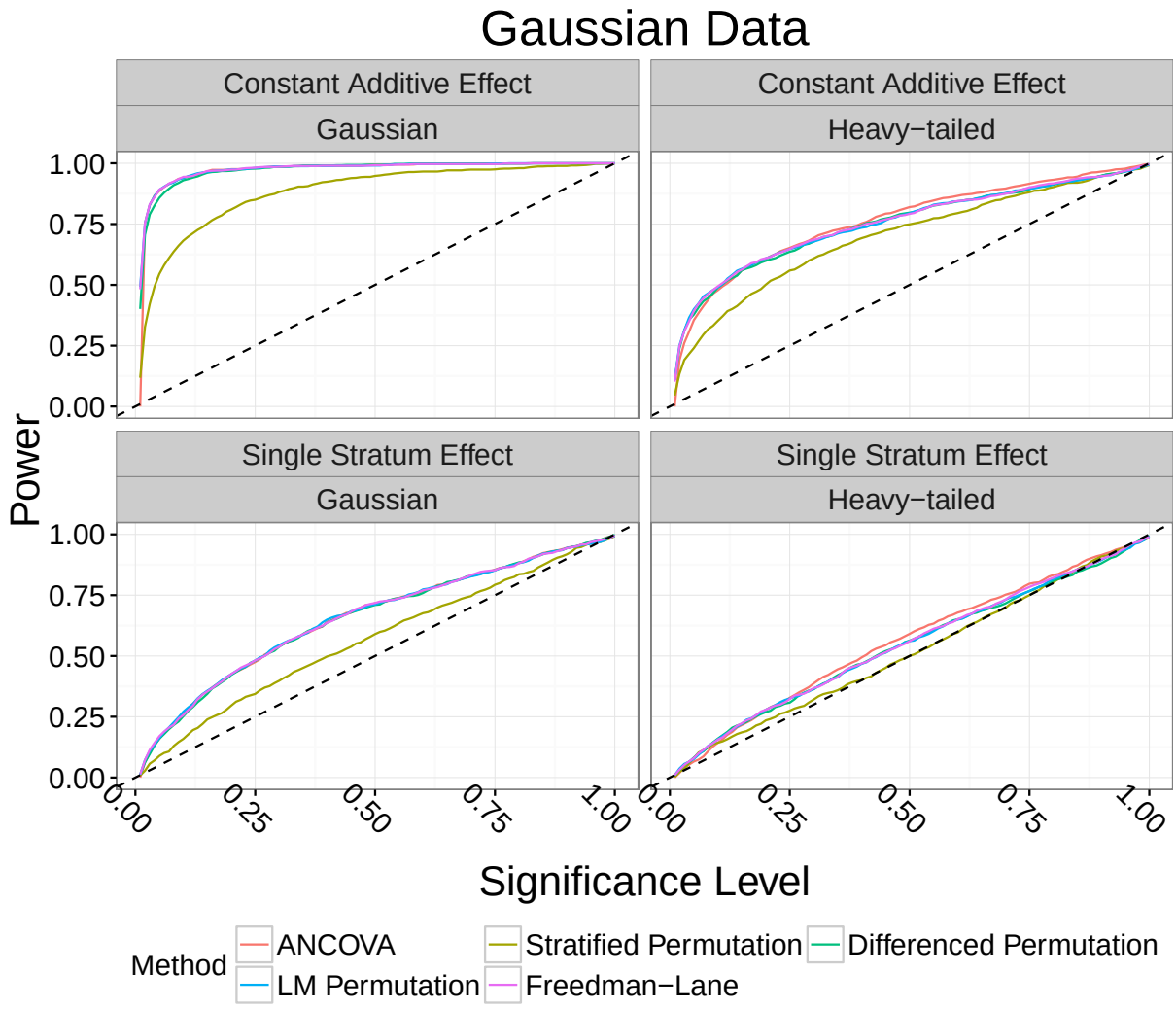

Figure 1: Empirical power curves for the Gaussian simulated data

the value of these power curves and the size of the test at level 0.05. In this case, the permutation tests (except for the simple stratified permutation) had slightly higher power than ANCOVA, but the difference was not substantial.

In our third set of simulations, we made the baseline measurements Y_{ij0} discrete and skewed. This type of data may occur in surveys where individuals are asked to rate their symptom severity on a discrete ordinal scale. We generated the baseline measures using independent draws from a Poisson distribution with mean 4 and censored values at 10. We conditioned on these observed baseline measures after generating them once. Then, for each of 1000 trials, we randomly assigned treatment to half of the individuals in each

Design		p-value				
Errors	Treatment Effect	ANCOVA	Stratified Permutation	Differenced Permutation	LM Permutation	Freedman-Lane
Gaussian	None	0.058	0.042	0.038	0.062	0.057
	Constant	0.886	0.545	0.858	0.891	0.892
	Single stratum	0.159	0.089	0.160	0.162	0.170
Heavy-tailed	None	0.032	0.040	0.045	0.042	0.040
	Constant	0.355	0.240	0.378	0.400	0.393
	Single stratum	0.066	0.080	0.082	0.076	0.080

Table 1: Empirical power at level 0.05 for Gaussian simulated data with balanced treatment groups.

stratum and generated errors taking on values 0.5 and -0.5 with equal probability. We constructed Y_{ij1} using Equation 2 with a treatment effect of $\gamma = 0.5$. To ensure that both baseline and outcome were integer scores, we let the Y_{ij1} be unobservable and defined the observed outcome as

$$\tilde{Y}_{ij1} = \begin{cases} 1 & \text{if } Y_{ij1} < 1 \\ 10 & \text{if } Y_{ij1} > 10 \\ \lfloor Y_{ij1} \rfloor & \text{otherwise} \end{cases}$$

The observed outcomes $\tilde{Y}_{ij1}$, the baseline measures Y_{ij0}, and the errors were discrete, violating normality assumptions.

We examined the effect of stratum size in this set of simulations. For each of the two treatment effect designs, we considered both equal and unequal stratum sizes. The simulations with equal stratum sizes were generated as before, where each stratum contained $n_j = 16$ individuals. The simulations with unequal stratum sizes were set up so the smallest stratum had only $n_1 = 8$ individuals, the next had $n_2 = 16$, and the largest had $n_3 = 24$. The design with a nonzero treatment effect in a single stratum and unequal stratum sizes was particularly unfavorable: the nonzero effect only occurred in the smallest stratum.

Figure 2 shows the estimated power curves in these four designs. When the treatment effect was constant, sample sizes within each stratum didn't matter; the two power curve

Design		p-value				
Errors	Treatment Effect	ANCOVA	Stratified Permutation	Differenced Permutation	LM Permutation	Freedman-Lane
Gaussian	None	0.046	0.031	0.028	0.042	0.045
	Constant	0.840	0.490	0.810	0.870	0.860
	Single stratum	0.120	0.050	0.090	0.100	0.120
Heavy-tailed	None	0.034	0.054	0.055	0.047	0.042
	Constant	0.230	0.200	0.270	0.260	0.270
	Single stratum	0.070	0.080	0.090	0.090	0.080

Table 2: Empirical power at level 0.05 for Gaussian simulated data with imbalanced treatment groups.

plots in the first row of Figure 2 look the same up to chance variation. When the treatment effect only occurred in one stratum, there was a substantial loss in power. It was nearly impossible to detect a difference in treatments when the affected stratum was too small. In practice, one does not usually know a priori which individuals will be affected by treatment (otherwise, those not affected would be excluded from the study altogether). This result suggests that when stratifying, one must be cautious not to stratify too finely, as it is impossible to measure an effect with an insufficient sample size.

All of the tests appeared to be conservative and to have lower than nominal level. As in the first set of simulations, the stratified permutation test using the raw outcomes had the lowest power of the five tests. The stratified permutation test using the change scores also had lower power than the three linear model tests. Presumably, this was because the differences $\tilde{Y}_{ij1} - Y_{ij0}$ were discrete and limited to a small number of values, so the test statistic could not vary greatly across permutations. Table 3 confirms this: power for the three linear model methods was high when the treatment was constant across strata, and power for all of the methods suffered when the effect was isolated in a single stratum.

It makes sense that in all of these simulations, the permutation test using the change scores as the response measure is more powerful than the permutation test using the raw outcomes. The baseline and outcome were highly correlated, so the change score had

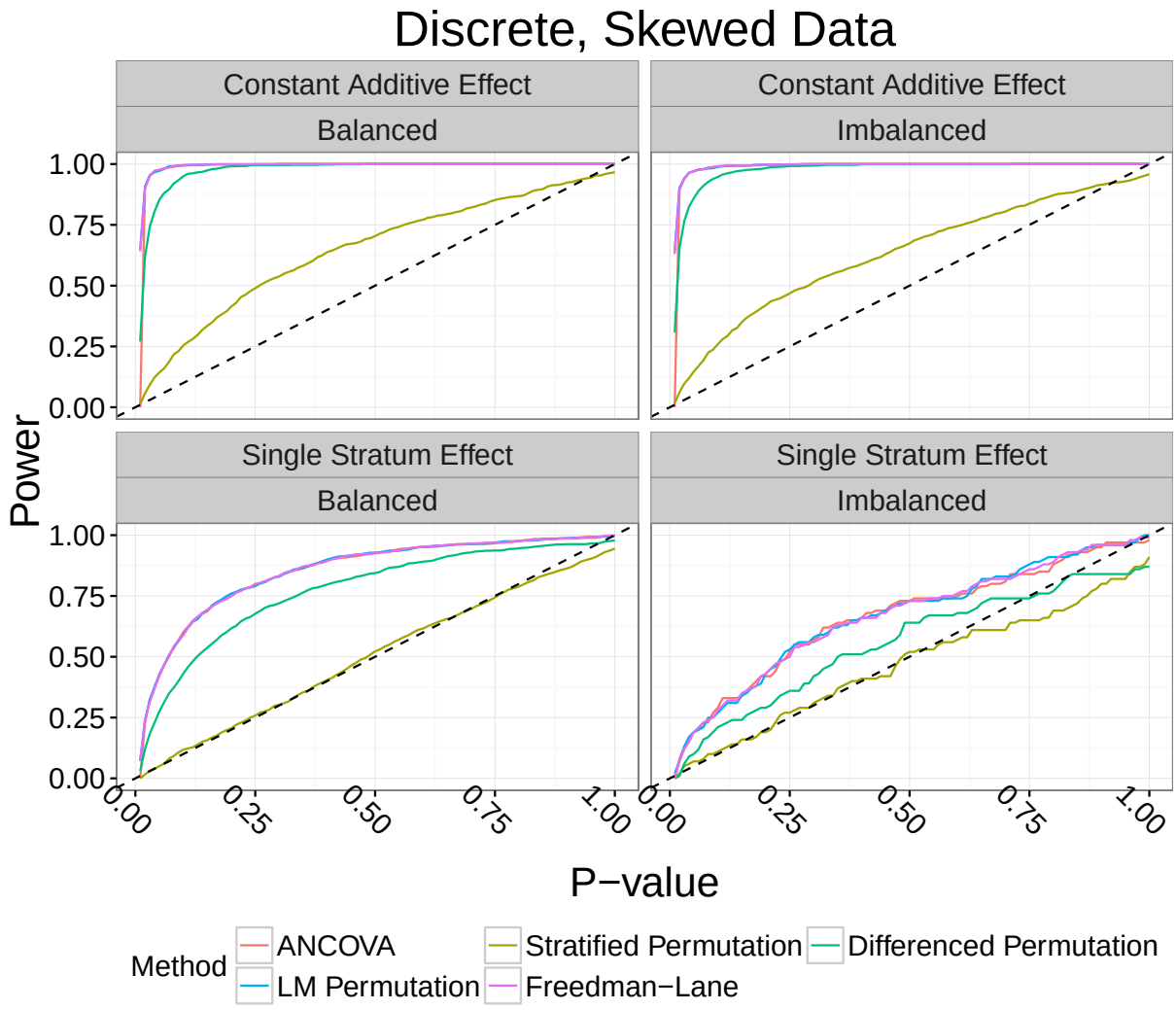

Figure 2: Empirical power curves for the skewed, discrete simulated data

Design		p-value				
Group sizes	Treatment Effect	ANCOVA	Stratified Permutation	Differenced Permutation	LM Permutation	Freedman-Lane
Balanced	None	0.036	0.025	0.021	0.035	0.037
	Constant	0.974	0.143	0.853	0.972	0.977
	Single stratum	0.422	0.050	0.275	0.429	0.424
Imbalanced	None	0.038	0.024	0.019	0.037	0.039
	Constant	0.971	0.147	0.860	0.970	0.972
	Single stratum	0.190	0.070	0.100	0.190	0.190

Table 3: Empirical power at level 0.05 for discrete, skewed simulated data

lower variance than the outcome alone. In additional simulations, we modified the data-generating process given by Equation 2 so that the outcome and baseline had a correlation of 0.25. In this case, the result was reversed: the differenced test had low power while the test using the raw outcomes had a power curve closer to the linear model-based tests. (See the supplementary Gaussian simulations for a more detailed treatment.) This behavior has been studied before by Frison and Pocock (1992), who recommended using the difference in outcome and baseline if their correlation is greater than 0.5 and to use the outcome only if their correlation is less than 0.5.

4 Clinical data results

We compared the parametric ANCOVA, the stratified permutation test, and the two linear model-based permutation tests using a dataset from a clinical trial comparing the effectiveness of two treatments for gastroesophageal reflux disease (GERD). A detailed discussion of the data and analysis is provided in a supplementary file. We summarize the analysis here. Patients were treated at eight sites in two different countries. At each site, patients were randomly assigned one of two treatments. Patients were observed for a week before receiving treatment and for a week after receiving treatment. On each of the fourteen days of observation, patients responded to a survey about their heartburn, regurgitation, and dyspepsia frequency and severity. These endpoints were measured on a discrete scale.

There were several additional endpoints, such as daily regurgitation, daily "hrdq", and daily dyspepsia, calculated from the survey measures. Daily "hrdq" was the primary endpoint. To reduce day-to-day variation, we averaged the measures from each week to obtain two observations per patient, one pre-treatment and one post-treatment.

We used site as the stratification variable, as this is the level at which treatment was randomized. We did not include country in the model, as a country-level effect should be captured in the site-level effects. The model used for the linear regression-based tests was defined as in Equation 1. This model allowed the intercept α_j to vary across sites and used the pretreatment, baseline measurement as the control variable X. The outcome and baseline had a low correlation (for instance, the correlation between pre- and post-treatment daily "hrdq" was 0.56 TO DO: WERE OTHER OUTCOMES EVEN LOWER? THE CUTOFF FOR CORRELATION WAS 0.5), so we used raw outcomes and not the change scores as the dependent variable. As our simulations and previous work (Frison and Pocock (1992)) show, when the correlation between baseline and outcome is low, it is more powerful to control for baseline outcomes using a model.

Figure 3 shows the distribution of each clinical endpoint for the two treatment groups. There is a clear difference in distributions for daily heartburn ("daily_heart"), daily "hrdq" ("daily_hrdq"), and heartburn frequency ("heart_freq"). The difference is less clear for daily regurgitation ("daily_regurg"), daily dyspepsia ("daily_dysp"), regurgitation frequency ("regurg_freq"), and dyspepsia frequency ("dysp_freq"). The distribution of outcomes for each endpoint is extremely right skewed, calling the assumption of normally distributed error terms into question.

Table 4 shows the p-values for the four tests and the seven continuous study endpoints. Overall, the results confirm our expectations based on visual comparison in Figure 3: one or more of the tests reject the null hypothesis that outcomes are the same between treatments for heartburn frequency, daily heartburn, and daily "hrdq," but not for any of the other endpoints. The p-values for the stratified, unadjusted permutation test have no consistent pattern: sometimes they are smaller than the p-values from the other tests and sometimes they are larger.

The three tests based on the linear model give qualitatively similar results here. The

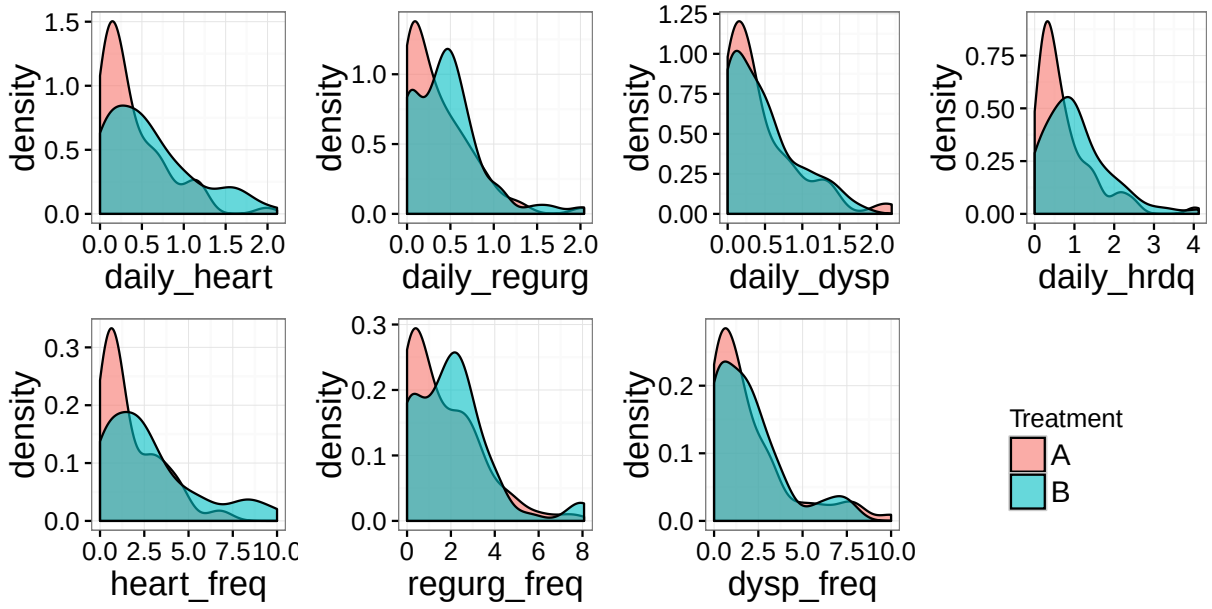

Figure 3: Distribution of each endpoint for the two treatments A and B

ANCOVA p-values tend to be smaller than the permutation linear model and residual permutation test p-values. The permutation tests are somewhat more conservative, as they give higher p-values in Table 4, accounting for the highly skewed outcome distributions. TO DO: IS THIS A CORRECT STATEMENT? CONSERVATIVE MEANS REJECT LESS OFTEN Our conclusions for the heartburn frequency, daily heartburn, and daily "hrdq" endpoints differ between ANCOVA and the permutation tests at significance level 0.05, but not at level 0.1. None of the three tests based on linear models is significant at level 0.1 for the other four endpoints; there is no endpoint which would be deemed significant using a permutation test but insignificant with ANCOVA. This suggests that the parametric test correctly discriminates between endpoints that are significantly different and endpoints that are not different between treatment A and treatment B.

5 Discussion

This paper adds to the literature comparing parametric and nonparametric tests. Our results match those of Vickers (2005) and Anderson and Legendre (1999), where parametric ANCOVA and regression-based t tests performed the same or better than the comparable

Endpoint	ANCOVA	Stratified Permutation	LM Permutation	Freedman-Lane
heart_freq	0.035	0.006	0.080	0.082
regurg_freq	0.136	0.118	0.280	0.220
dysp_freq	0.565	0.948	0.616	0.592
daily_heart	0.032	0.004	0.056	0.068
daily_regur	0.142	0.174	0.286	0.246
daily_hrdq	0.043	0.012	0.088	0.098
daily_dysp	0.582	0.810	0.756	0.722

Table 4: Comparison of p-values from four tests, for each continuous endpoint.

nonparametric tests. We simulated a variety of data-generating processes, ranging from the ideal case when all assumptions are met, to the case when data are non-normal, discrete, or skewed. In the supplementary simulation files, we also considered data with other possible issues such as heterogeneous treatment effects and unequal sample sizes. The linear regression-based tests, both the parametric ANCOVA and the permutation tests, suffered a loss of power when the ANCOVA assumptions were violated. However, the tests remained comparable to each other. It is a matter of taste which test one chooses for their experiment: while the parametric test may be robust to violations of its assumptions, it is somewhat reassuring that the permutation test can exactly match the randomization that was done while making no distributional assumptions. Applying a linear model based permutation test as a secondary analysis can give insight into how ANCOVA results depend on the method's assumptions (making it conservative, anti-conservative, or not systematically affecting the p-value in a particular direction), and give evidence that inferences based on ANCOVA are reliable.

Permutation tests do not come entirely free of assumptions, though: subjects must be exchangeable if we generate the permutation distribution assuming that all possible allocations of treatment which preserve the number of treated units are equally likely. Romano (1990) warns against using permutation tests naively if items are not truly exchangeable. For instance, he points to the case where observations have unequal variances. This is a

problem in practice as one cannot observe errors; it is a leap of faith to assert that they are homogeneous and therefore exchangeable. Boik (1987) illustrates this phenomenon using the traditional F test and its permutation counterpart, and demonstrates by simulation that the latter has larger than nominal level when the error variances are unequal. Randomized experiments mitigate this problem: by definition, treatment assignment is statistically independent of all other variables (possibly conditioning on strata).

It is important to note that the permutation tests described here are exact only in the context of randomized experiments. Treatment is assigned at random and is therefore statistically independent of the covariates X and the errors ε. In observational studies, treatment may be associated with X, ε, or both, often in a way that is difficult or impossible to model. The statistical independence guaranteed by randomized experiments enables us to construct permutation distributions while holding X fixed. When exchangeability doesn't hold, we cannot disentangle the effect of Z from the effect of X. One must provide evidence that variables in observational studies are exchangeable in order to achieve an approximately exact test. The onus is on the researcher to make the case that the variable being permuted is uncorrelated with the other variables being held fixed. This can be done visually using scatterplots and residual plots (see e.g. Freedman and Lane (1983)) and from knowledge of how the data arose and were collected.

One must be mindful of which covariates to include as controls. Ideally, all covariates related to the outcome are known and observed. In that case, the coefficient for treatment in a fully saturated linear model (including all covariates and their interactions) is an asymptotically consistent estimate of the average treatment effect (Lin (2013)). It is rare that one knows *all* relevant covariates in practice; failure to control for all relevant variables can introduce bias in some permutation tests. Gail et al. (1988) propose a randomization test based on residuals from an exponential family model and find that omitting relevant covariates leads to tests with higher than nominal level.

The method of controlling for baseline covariates matters, too. The naive way is to consider change scores, subtracting baseline measures from the outcome and doing hypothesis testing using the change as the dependent variable. Our simulations confirm the suggestion of Frison and Pocock (1992) to use change scores only when there is a strong correlation

between baseline and outcome. Weak correlations between baseline and outcome occur often in practice, as was the case with our GERD dataset. When the correlation is weak, the test of change scores may be less powerful than ignoring the baseline altogether. Instead, we suggest incorporating the baseline in a regression model. This is more general than using differences; treating the change scores as the dependent variable is a special case of the linear regression that constrains the coefficient of the baseline measure to be 1. The regression approach is more flexible and demonstrably more powerful.

SUPPLEMENTARY MATERIAL

All files are also available at `https://github.com/kellieotto/ancova-permutations`.

Gaussian data simulations: Simulations using Gaussian data in Section 3, including code and results. (PDF)

Imbalanced design simulations: Simulations using data with different stratum sizes in Section 3, including code and results. (PDF)

Skewed data simulations: Simulations using discrete, skewed data in Section 3, including code and results. (PDF)

Clinical trial data: Dataset used in Section 4 to compare methods. (csv)

README: Data descriptor file. (txt)

Results: Detailed explanation, code, and results of comparing methods using the clinical trial dataset in Section 4. (PDF)

References

Anderson, M. J. and P. Legendre (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation 62*(3), 271–303.

Anderson, M. J. and J. Robinson (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics 43*(1), 75–88.

Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology 40*(1), 26–42.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Prentice-Hall.

de Winter, D. and J. C. F. Dodou (2010). Five-point likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation 15*(11), 1–12.

Fisher, R. A. (1935). *Design of Experiments*. New York: Hafner.

Freedman, D. and D. Lane (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics 1*(4), 292–298.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Frison, L. and S. J. Pocock (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine 11*(13), 1685–1704.

Gail, M. H., W. Y. Tan, and S. Piantadosi (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika 75*(1), 57–64.

Glass, G. V., P. D. Peckham, and J. R. Sanders (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research 42*(3), 237–288.

Kennedy, F. E. (1995). Randomization Tests in Econometrics. *Journal of Business & Economic Statistics 13*(1), 85–94.

Kennedy, P. E. and B. S. Cade (1996). Randomization tests for multiple regression. *Communications in Statistics - Simulation and Computation 25*(4), 923–936.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco, California: Holden-Day.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics 7*(1), 295–318.

Ludbrook, J. and H. Dudley (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician 52*(2), 127–132.

Manly, B. F. J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition*. CRC Press.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science 5*(4), 465–472. translated by Dabrowska, D M and Speed, T P.

Pesarin, F. and L. Salmaso (2010). *Permutation tests for complex data: theory, applications and software.* John Wiley & Sons, Inc.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society 4*(1), 119–130.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika 29*(3/4), 322–335.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association 85*(411), 686–692.

Still, A. W. and A. P. White (1981). The approximate randomization test as an alternative to the F test in analysis of variance. *British Journal of Mathematical and Statistical Psychology 34*(2), 243–252.

Vexler, A., A. D. Hutson, and X. Chen (2016). *Statistical testing strategies in the health sciences*. CRC Press.

Vickers, A. J. (2005). Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology 5*, 35.

Winkler, A. M., G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols (2014). Permutation inference for the general linear model. *NeuroImage 92*, 381–397.

Zimmerman, D. W. (2006). Two separate effects of variance heterogeneity on the validity and power of significance tests of location. *Statistical Methodology 3*(4), 351–374.