University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2008 *Paper* 219

Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models

Michael Rosenblum[*] Mark J. van der Laan[†]

[*]Division of Biostatistics, School of Public Health, University of California, Berkeley, mrosenbl@jhsph.edu

[†]Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models

Michael Rosenblum and Mark J. van der Laan

Abstract

Regression models are often used to test for cause-effect relationships from data collected in randomized trials or experiments. This practice has deservedly come under heavy scrutiny, since commonly used models such as linear and logistic regression will often not capture the actual relationships between variables, and incorrectly specified models potentially lead to incorrect conclusions. In this paper, we focus on hypothesis test of whether the treatment given in a randomized trial has any effect on the mean of the primary outcome, within strata of baseline variables such as age, sex, and health status. Our primary concern is ensuring that such hypothesis tests have correct Type I error for large samples. Our main result is that for a surprisingly large class of commonly used regression models, standard regression-based hypothesis tests (but using robust variance estimators) are guaranteed to have correct Type I error for large samples, even when the models are incorrectly specified. To the best of our knowledge, this robustness of such model-based hypothesis tests to incorrectly specified models was previously unknown for Poisson regression models and for other commonly used models we consider. Our results have practical implications for understanding the reliability of commonly used, model-based tests for analyzing randomized trials.

1 Introduction

Data sets from randomized, clinical trials are often analyzed using models such as linear regression, logistic regression, or Poisson regression models. The validity of conclusions drawn from model-based analyses generally relies on the assumption that the model is correctly specified, that is, the assumption that the statistical model accurately represents the true data generating distribution. Robins (1994, 2004), Freedman (1997, 2008a), and Berk (2004), among others, have drawn much needed attention to the fact that when this assumption is false it may lead to false conclusions. Furthermore, in medical studies and studies involving biological systems in general, due to the complexity of relationships between variables, simple regression models may fail to accurately represent the true relationships between these variables. It may not even be possible to detect when a model is incorrectly specified, since for the sample sizes available in many applications, diagnostics of model fit have good power to detect only a limited number of the potential ways that a model may fail to be correctly specified (Freedman, 2005). It is therefore important to understand when reported results based on regression models will be reliable, even when the models being used are incorrectly specified.

In this paper, we examine the properties of incorrectly specified regression models (also called misspecified models) when they are used in hypothesis tests in randomized trials. The null hypothesis we consider throughout the paper is that the treatment being evaluated has no effect on mean outcome within subpopulations defined by a given set of baseline variables. For example, in a randomized trial of an HIV vaccine, this null hypothesis could be that the vaccine has no effect on HIV infection rates for subpopulations defined by age, study site, and presence of other sexually transmitted infections measured at baseline. A standard technique for testing such a hypothesis involves fitting a regression model for the mean outcome given the treatment and baseline variables. The null hypothesis is rejected at level 0.05 if the 95%-confidence interval for the coefficient corresponding to the treatment variable in this model excludes 0. When the model is correctly specified, such a test has asymptotically correct Type I error, meaning that the probability of rejecting the null hypothesis when it is in fact true converges to a value at most 0.05 as sample size goes to infinity.

However, as argued by Robins (2004), for some classes of models, when the regression model is incorrectly specified, Type I error may be quite large even for large sample sizes. It has been an open problem to determine which models have this problem, and which models are guaranteed to have asymptotically correct Type I error, even when these models are misspecified.

Our main contribution is showing for a surprisingly large class of commonly used regression models, that standard hypothesis tests based on these models (but using robust variance estimators) are protected from the above problem. That is, for this large class of regression models, the corresponding hypothesis tests are guaranteed to have asymptotically correct Type I error, regardless of whether the actual data generating distribution behaves according to the model or not. This is a non-trivial result, since the regression models we consider will in general be incorrectly specified even under the null hypothesis of no mean treatment effect within strata of baseline variables. Our results enable the use of the model-based tests described in Section 4, without fear of inflated Type I error, at least asymptotically, under the assumptions given in Section 3.

Examples of this robustness to misspecified models, for some types of linear regression models, have been shown in (Robins, 2004; Freedman, 2008a). We show that this property of robustness to incorrectly specified models holds for two very large classes of commonly used models. First, we show it holds for a large class of linear regression models and give a simple procedure for augmenting any linear model to ensure it has this robustness property. Second, we show the robustness property holds for many commonly used models including logistic regression models, probit regression models, binary regression models with complementary log-log link, and Poisson regression models; the main requirement is that the linear part in such models be of a certain commonly used form that we describe in Section 5.

Important work has been done using semiparametric methods to construct estimators and hypothesis tests that are robust to incorrectly specified models in the setting of randomized trials, for example (Robins, 1986, 1994; van der Laan and Robins, 2002; Tsiatis et al., 2008; Zhang et al., 2008; Moore and van der Laan, 2007; Rubin and van der Laan, 2007). The emphasis of this paper, in contrast, is on standard regression-based methods. Regression-based methods have the advantage that they may be more familiar to many statisticians, who already have much expertise implementing regression analyses in statistical software.

Throughout the paper, we assume the true data generating distribution, which is unknown to the experimenter, may not be in the experimenter's model. We only make the following three assumptions on the data generating distribution: treatment is randomly assigned, all variables are bounded, and each subject's data is i.i.d. from an unknown distribution. The first assumption will be true in all randomized trials. The second holds for most variables of interest. The latter assumption, also made in (Robins, 2004; Tsiatis et al., 2008; Zhang et al., 2008; Moore and van der Laan, 2007), will only be true in some types of randomized trials; we explain this further in the discussion section. The focus of this paper is hypothesis testing, and we note that our robustness results do not, in general, imply analogous results for estimation; we discuss this issue in Section 8.

In the next section, we give an example that typifies how regression models are used to draw conclusions from randomized trials, and we show how our results apply to this example. We describe the hypothesis testing problem and robustness property considered in this paper in Section 3. In Sections 4 and 5 we give our main results in terms of robustness of certain regression-based hypothesis tests to incorrectly specified models. In Sections 6 and 7, we use simulations and a data example from a recently completed randomized trial (Padian et al., 2007) to compare the performance of regression-based hypothesis tests to other robust methods. We discuss the practical implications of our results in Section 8 and defer all proofs to the Web Appendices. (The URL for the Web Appendices is given at the end of the paper.)

2 Example of a Regression-Based Hypothesis Test in a Randomized Trial

To illustrate how our results are motivated by issues arising in the analysis of clinical trials, we consider the recently completed "Randomized Trial of Inhaled Cyclosporine in Lung-Transplant Recipients" (Iacono et al., 2006). The treatment was an inhaled drug to help prevent rejection after lung transplantation. Half the subjects were randomly assigned a placebo drug. The primary outcome was the number of severe (grade 2 or higher) rejection events per year of follow-up time. We refer to this count using the variable REJECTIONS. In one of the main analyses, a Poisson regression model was used to test for differences in mean outcome between the treatment group and control group, within subpopulations defined by several baseline variables. These baseline variables included indicators of whether there was a serologic mismatch between donor and recipient (denoted by V_1), and of whether a rejection episode had occurred before the first inhaled treatment was given (denoted by V_2). We denote the treatment group by $A = 1$ and the control group by $A = 0$. The Poisson model specifies that the logarithm of the conditional mean of the variable REJECTIONS given treatment and baseline variables has the form:

$$\log \mathrm{E}(\mathrm{REJECTIONS} \ \mid A, V_1, V_2) = \beta_0 + \beta_1 A + \beta_2 V_1 + \beta_3 V_2. \tag{1}$$

(Throughout the paper, log refers to the natural logarithm.) The model also specifies that conditioned on A, V_1, V_2, the variable REJECTIONS has a Poisson distribution.

This Poisson model was used to carry out a hypothesis test of whether there was any mean treatment effect within strata of baseline variables V_1, V_2. First, the model was fit using maximum likelihood estimation, giving an estimate of the coefficient vector β. Then, if the 95% confidence interval around the estimate for the coefficient β_1 were found to exclude 0, it would be concluded that there was a statistically significant difference in the outcome due to the treatment (inhaled cyclosporine). (In the case of this trial, the observed difference turned out not to be statistically significant.)

Standard arguments to justify the validity (in terms of Type I error) of such model-based hypothesis tests rely on the model being correctly specified, at least approximately. In the above example, there is no *a priori* reason to think this should be the case, and no reason based on subject-knowledge about rejection events in lung transplant recipients is given in the paper. Thus, it may be the case that the above model is not correctly specified. Our goal in presenting this example is not to criticize a particular analysis. We merely wanted to illustrate the commonly used practice of model-based hypothesis tests in randomized trials, and point out the potential for models to be misspecified. Freedman (2008a) provides many other examples of randomized trials in which regression models are used.

Our main result provides asymptotic guarantees for Type I error without having to assume the model is correctly specified. In particular, our results imply that the above hypothesis test will have asymptotically correct Type I error, if the confidence interval is instead computed using the sandwich estimator of Huber (1967) (described in detail in Web Appendix B), and if the subjects represent a random sample from a larger population. In this case, with probability tending to 0.95, for large sample sizes, one is protected against falsely concluding there is a mean effect of the treatment within some stratum of the baseline variables V_1 and V_2, when no such effect exists, even when the model used is misspecified.

3 Notation, Assumptions, Hypothesis Testing Problem, and Robustness Property

We now explain the underlying assumptions and goal of the hypothesis testing problem that we address in the remainder of the paper. We start by introducing our notation. Let Y represent the outcome of interest, A represent the treatment assignment, and V represent a vector of baseline variables such as age, sex, and past health status. We assume there are k different treatments being evaluated in the randomized trial, so that A takes values in $\{0, \ldots, k-1\}$; many randomized trials have $k = 2$, corresponding to a single treatment ($A = 1$) being compared to a control ($A = 0$). We consider regression models $m(A, V|\beta)$ of the mean outcome given treatment and baseline variables: $E(Y|A, V)$. Some of these models (e.g. Poisson regression) specify additional characteristics of the distribution of Y. As a technical condition, we assume all variables are bounded[1]. We also note that all of our results hold even when the regression model used (such as a Poisson regression model) assumes that variables are unbounded. This is because, as discussed below, our results hold whether or not the assumptions underlying the regression model are true. Our results, however, only guarantee asymptotically correct Type I error, and we note that when the regression model used is a poor approximation to the true data generating distribution, this may result in low power, as discussed in Section 6.

Since we never assume that the model $m(A, V|\beta)$ is correctly specified, it may be better to think of it as a "working model"; that is, since we never assume the data generating distribution obeys the constraints of the model, $m(A, V|\beta)$ can be considered merely as a mathematical formula given as input, along with the data, to a hypothesis testing procedure. For example, the Poisson model (1) in the previous section can be viewed as a working model, that is, simply a formula used by statistical software to generate a 95%-confidence interval for the coefficient β_1; this confidence interval is used to decide whether to reject the null hypothesis or not, depending on whether it excludes 0. The purpose of this paper is to prove guarantees for the Type I error of such hypothesis testing procedures, without assuming the model is correctly specified.

The only assumptions we make on the data generating distribution are that each subject's data is i.i.d. from an unknown distribution (which is a common assumption in the superpopulation inference framework, further discussed in Web Appendix F), that all variables are bounded, and that treatments are randomly assigned. We also prove our results under a modified set of assumptions that better represents the actual way in which data is generated in randomized trials; in particular, these modified assumptions allow for treatment being randomly assigned to fixed proportions of the study subjects, instead of being assigned i.i.d. This modified set of assumptions is given in Web Appendix D.

Our focus throughout the paper is testing the null hypothesis of no mean treatment effect within strata of a set of baseline variables V. More formally, we define our null hypothesis as follows:

Null Hypothesis: For all treatments a_1, a_2,

$$E(Y|A = a_1, V) = E(Y|A = a_2, V). \tag{2}$$

The above expectations are taken with respect to the true data generating distribution.

[1] Boundedness of variables is used in our proofs to establish integrability of the log-likelihood and its first two derivatives. We find the boundedness assumption natural in that most variables in practice will have minimum and maximum possible values.

Because we assume the data come from a randomized trial, $E(Y|A = a, V = v)$ has a causal interpretation as the mean outcome that would have been observed had everyone with baseline variables $V = v$ in the population from which the trial participants were drawn been assigned treatment $A = a$. Cast in this light, our null hypothesis is that the treatment has no effect on the mean outcome, within strata of baseline variables.

Note that this null hypothesis is weaker than the null hypothesis of *no effect at all* of treatment on the distribution of the outcome within such strata[2]; our null hypothesis only posits that the *mean* of the outcome within such strata is not affected by which treatment is administered. However, in the special case that Y is binary, so that the conditional mean of Y characterizes the entire conditional distribution of Y, our null hypothesis simplifies and is equivalent to no effect at all of treatment on the distribution of the outcome within strata of V. In this case, certain permutation tests can also be used to test our null hypothesis; we will compare the power of our regression-based tests to a permutation test of Rosenbaum (2002) in Section 6. Throughout this paper, we are concerned with testing the null hypothesis (2); however, in Web Appendix C we also prove a result for testing a different type of null hypothesis–that of no effect modification by baseline variables.

The property we will prove for hypothesis tests based on many commonly used regression models is the following:

Robustness Property: (3)

We say a hypothesis test at level α is asymptotically robust to misspecification if for any data generating distribution satisfying the above null hypothesis (2), the asymptotic probability of rejecting the null hypothesis is at most α. If a hypothesis test satisfies this property at all levels α, we simply say it is asymptotically robust to misspecification.

[2] In our framework, we say there is no effect at all of treatment on the distribution of the outcome within strata of V if the treatment A is mutually independent of baseline variables V and outcome Y.

4 Main Result

Our main result is that for a large class of commonly used regression models, a type of hypothesis test (that we describe in detail below) based on such models has the robustness property (3). We make the following assumptions:

(A1) The data are generated as described in Section 3.

(A2) A regression model $m(A, V|\beta)$ in one of the classes given in Section 5 below is used. For example, if the outcome is dichotomous, one could use the logistic regression model

$$m(A, V|\beta) = \text{logit}^{-1}(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV). \tag{4}$$

(A3) β_i is a pre-specified coefficient of a term containing the treatment variable A in the linear part of the model chosen in (A2). For example, if using the logistic model (4), one could specify either β_1 or β_3. One can also use more than one of these coefficients. For example, one could use both β_1 and β_3 if using the logistic model (4).

Consider the following hypothesis test:

Hypothesis Test: (*)

For concreteness, we consider the case of testing at level $\alpha = 0.05$. The parameter β is estimated with ordinary least squares estimation if the model used is linear; otherwise it is estimated with maximum likelihood estimation. The standard error is estimated by the sandwich estimator, which can easily be computed with standard statistical software; we describe the sandwich estimator in detail in Web Appendix B. If a single coefficient β_i is chosen in (A3), then the null hypothesis of no mean treatment effect within strata of V is rejected at level 0.05 if the estimate for β_i is more than 1.96 standard errors from 0. If several coefficients are chosen in (A3), one can perform a similar test based on a Wald statistic that uses the estimates of these coefficients along with their covariance matrix based on the sandwich estimator; we describe this procedure in Web Appendix B.

We note that in some cases, such as when the design matrix is not full rank or the maximum likelihood estimator fails to converge, the estimators we consider will be undefined. *We therefore specify that regardless of whether the estimate for the coefficient β_i is more than 1.96 standard errors from 0, we always fail to reject the null hypothesis if the design matrix has less than full rank or if the maximum likelihood algorithm fails to converge.* Since standard statistical software (e.g. Stata or R) will return a warning message when the design matrix is not full rank or when the maximum likelihood algorithm fails to converge, this condition is easy to check.

The main result of this paper is the following theorem (proved in Web Appendix D):

Theorem: *Under assumptions (A1)-(A3), the hypothesis test (*) has the robustness property (3). That is, it has asymptotic Type I error at most* 0.05*, even when the model is misspecified.*

We point out that the above theorem is non-trivial, since the regression models in the two classes we describe in Section 5 below will generally be incorrectly specified if V is high-dimensional, under the null hypothesis (2). This follows since even under this null hypothesis of $E(Y|A,V)=E(Y|V)$, a correct model of $E(Y|V)$ would have to exactly capture how the mean of the outcome Y depends on the baseline variables V; when V is high-dimensional, this is generally impossible unless the mechanisms that determine Y have simple, well understood functional forms.

We briefly outline the main steps in the proof of the above theorem; readers who are mainly interested in the application of this method in practice may prefer to skip to Sections 5-7. The full proof of the above theorem is given in Web Appendix D. The main work of the proof is showing that under the assumptions (A1)-(A3) above, the estimate $\hat{\beta}_i$ of the coefficient β_i is asymptotically normal, and converges to 0 under the null hypothesis (2). Once this is proved, the theorem follows from the fact that robust variance estimates computed by the sandwich estimator are asymptotically correct even for misspecified models. That $\hat{\beta}_i$ is asymptotically normal follows from a standard result characterizing the convergence of maximum likelihood estimators for generalized linear models, as given in Theorem 5.23 in (van der Vaart, 1998, pg. 53). It remains, then, to show that under the null hypothesis (2), $\hat{\beta}_i$ converges to 0 as sample size goes to infinity. The proof of this fact is the main technical contribution of this paper. It relies on A being independent of V (as is the case in a randomized trial), on the null hypothesis (2), and on the exponential form of the likelihood for generalized linear models.

5 Classes of Regression Models that Guarantee the Robustness Property

We now describe two classes of regression models that can be used in the hypothesis test described in the previous section; when any of these models is used, the resulting hypothesis test is guaranteed to have the robustness property (3). We emphasize that these models are considered ”working models,” in that we never assume they are correctly specified. We first give a class of linear models and then give a class of generalized linear models. The choice of model can have a large effect on the power of the hypothesis test, which we explore in Section 6 below and in Web Appendix A. Note that the choice of which model to use must be made prior to looking at any of the data from the randomized trial. Otherwise, the risk of such data snooping would be that Type I error could be increased.

5.1 Linear Models

We exactly characterize the class of linear models for which the hypothesis test (*) has the robustness property (3).

Before giving a formal characterization of this class in (5) below, we give an informal description and several examples. Consider the special case of A a binary treatment, taking values 0 and 1. Roughly speaking, a linear model is in our class if for every term $f(A, V)$ in the model, the terms $f(1, V)$ and $f(0, V)$ are contained in the model as well, or else these corresponding terms must be linear combinations of other terms in the model. For example, the linear model

$$m_1(A, V|\beta) = \beta_0 + \beta_1 A + \beta_2 V,$$

is in our class, since corresponding to the A term, we also have an intercept term. Also, the following models are in our class:

$$m_2(A, V|\beta) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV,$$

$$m_3(A, V|\beta) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV + \beta_4 V^2 + \beta_5 AV^2,$$

since for the term AV we have the corresponding main term V, and for the term AV^2, we have corresponding term V^2.

The following model

$$m_4(A, V|\beta) = \beta_0 + \beta_1 e^V + \beta_2 e^{(2A-1)V},$$

does not contain a term (or linear combination of terms) corresponding to setting $A = 0$ in $e^{(2A-1)V}$, is not in our class of models, and does not have the robustness property (3). However, if we set $A = 0$ in $e^{(2A-1)V}$, producing the term e^{-V}, and add this term to the above model, we get the following model that is in our class:

$$m_5(A, V|\beta) = \beta_0 + \beta_1 e^V + \beta_2 e^{-V} + \beta_3 e^{(2A-1)V},$$

and so has the robustness property (3). This illustrates a general method by which one can add terms to any existing linear model to obtain a model with the robustness property (3).

In the above models m_1, m_2, m_3, all terms containing the treatment variable A are a product of a function of A and a function of the baseline variables V. The class of linear models of this form, in which for each such product term $f(A)g(V)$ the model also contains the term $g(V)$, was shown by Robins (2004) to have the robustness property (3). Our result extends Robins' class to a larger class of linear models which we formally define next. This larger class includes model m_5 for example. We show in Web Appendix D that this larger class is the largest possible, in that it contains all linear models having robustness property (3).

We now formally define our class of linear models to be all models of the form:

$$m(A, V|\beta) = \sum_j \beta_j^{(0)} f_j(A, V) + \sum_k \beta_k^{(1)} g_k(V), \tag{5}$$

where $\{f_j, g_k\}$ can be any functions bounded on compact sets such that for each term $f_j(A, V)$, we have $E(f_j(A, V)|V)$ is a linear combination of terms $\{g_k(V)\}$. We denote the parameter vector $(\beta^{(0)}, \beta^{(1)})$ simply by β. Since our setting is a randomized trial in which the probabilities of treatment assignment A are independent of baseline variables V and are set by the experimenters, one can always directly compute the conditional expectations $E(f_j(A, V)|V) = \sum_a f_j(a, V)p(a)$.

The theorem in Section 4 states that when the hypothesis test (*) uses a linear model of type (5), it has the robustness property (3), which guarantees asymptotically correct Type I error for testing the null hypothesis (2) even when the model is misspecified. The converse is also true. That is, when the hypothesis test (*) uses a linear model not having the property described just after (5), but for which the terms are linearly independent, then it will not have the robustness property (3). This is proved in Web Appendix D.

5.2 Generalized Linear Models

In this section we define our class of generalized linear models for which the hypothesis test (*) has the robustness property (3). Before giving a formal characterization of our class of generalized linear models in (6) below, we give an informal description and some examples. Consider the following types of generalized linear models: logistic regression, probit regression, binary regression with complementary log-log link function, and Poisson regression with log link function. A generalized linear model is in our class if it is of one of these types, and if the linear part is of a commonly used form defined precisely in (6) below. We now give specific examples of generalized linear models in this class. We show just a few of the many possibilities. Note that the linear parts (for example, $\beta_0 + \beta_1 A + \beta_2 V$) shown in any of the models below can be used in any of the other models.

Examples of Generalized Linear Models for which Robustness Property 3 Holds:

1. Logistic Regression: For Y binary and $\text{logit}(x) = \log(x/(1-x))$, the following model for $P(Y = 1|A, V)$:
$$m_6(A, V|\beta) = \text{ logit}^{-1}(\beta_0 + \beta_1 A + \beta_2 V),$$
2. Probit Regression: For Y binary and $\Phi(x)$ the cumulative distribution function of the standard normal, the following model for $P(Y = 1|A, V)$:
$$m_7(A, V|\beta) = \Phi(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV).$$
3. Binary Regression with complementary log-log link function: The complementary log-log function is $\zeta(\mu) = \log(-\log(1-\mu))$. The following model for $P(Y = 1|A, V)$:
$$m_8(A, V|\beta) = \zeta^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V^2 + \beta_3 AV^2\right).$$
4. Poisson Regression: For Y a "count" (that is, Y a nonnegative integer), the Poisson (log-linear) model:
$$\log m_9(A, V|\beta) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV.$$

We now give a formal description of the generalized linear models in our class. Consider the following types of generalized linear models[3]: logistic regression, probit regression, binary regression with complementary log-log link function, and Poisson regression with log link function. We define our class of generalized linear models to be any generalized linear model from the previous list, coupled with a linear part of the following form:

$$\eta(A, V|\beta) = \sum_j \beta_j^{(0)} f_j(A) g_j(V) + \sum_k \beta_k^{(1)} h_k(V), \tag{6}$$

for any measurable functions $\{f_j, g_j, h_k\}$ such that for all j, there is some k for which $g_j(V) = h_k(V)$; we also assume the functions $\{g_j, h_k\}$ are bounded on

[3] See McCullagh and Nelder (1998) for more details on generalized linear models.

compact subsets of $\mathbb{R}^q$, where V has dimension q. Note that (6) is more restrictive than the constraint (5) on linear models above, but includes many models used in practice (such as the models given as examples above). A model being in this class is a sufficient condition, but not a necessary condition, for the hypothesis test (*) to have the robustness property (3).

Our results also hold for other families of generalized linear models, such as Gamma and Inverse-Gaussian models with canonical link functions, but certain regularity conditions[4] are needed for the robustness property (3) to hold in these cases. For the models described above, no regularity conditions beyond those given just after (5) and (6) are needed.

6 Simulation Studies

We use simulations to compare the power of the regression-based method of this paper to other hypothesis testing methods. Model-based hypothesis tests have been shown to sometimes have more power than the intention-to-treat based hypothesis test (Robinson and Jewell, 1991; Hernández et al., 2004; Moore and van der Laan, 2007). However, depending on the data generating distribution and working model used, model-based hypothesis tests can also have lower power than the intention-to-treat based test. In this section, we examine the power of six robust methods under various data generating distributions. We only present the results of one set of simulations that are representative of the findings from a larger set of simulations; this larger set of simulations is given in Web Appendix A.

For simplicity, we consider randomized experiments with binary outcome Y, two possible treatments $A = 0$ and $A = 1$, and a continuous-valued baseline variable V. The hypothesis being tested is that of no mean treatment effect within strata of this baseline variable, as formally defined in (2). Since the robustness property (3) is the main focus of this paper, we chose to compare the method of this paper to other methods that also have this robustness property. Below, we give summaries of the methods we will be comparing, each of which has the robustness property (3); detailed descriptions are given in Web Appendix A.

Hypothesis Testing Methods:

M0: Regression-based test: This is the hypothesis testing method (*) described in Section 4. The estimated coefficients corresponding to all terms

[4] These regularity conditions are primarily technical, and result from the fact that for Gamma and Inverse-Gaussian models, the log-likelihood is only defined when the linear part $\eta(A, V|\beta)$ in (6) above is strictly positive.

in the working model that contain the treatment variable are combined into a Wald statistic, as described in Web Appendix B.

M1: Intention-to-treat based test: Estimate the risk difference by taking the difference between the empirical means of the two treatment groups. Reject the null hypothesis whenever the 95% confidence interval for the estimated risk difference excludes 0.

M2: Cochran-Mantel-Haenszel test: (Cochran, 1954; Mantel and Haenszel, 1959) First, the baseline variable is discretized. Then, the Cochran-Mantel-Haenszel test is run.

M3: Permutation test: (Rosenbaum, 2002) First, the binary outcome Y is regressed on the baseline variable V using logistic regression. Pearson residuals for each observation are calculated based on the model fit. Then, the residuals for observations in which $A = 1$ are compared to those for $A = 0$ using the Wilcoxon rank sum test.

M4: Targeted Maximum Likelihood based test: (Moore and van der Laan, 2007; van der Laan and Rubin, 2006) The risk difference is estimated, adjusting for the baseline variable using the targeted maximum likelihood approach; the null hypothesis is rejected if the 95% confidence interval for the risk difference excludes 0.

M5: Augmented Estimating Function based test: (Tsiatis et al., 2008; Zhang et al., 2008) The log odds ratio is estimated, using an estimating function that is augmented to adjust for the baseline variable; the null hypothesis is rejected if the 95% confidence interval for the log odds ratio excludes 0.

In all the scenarios we consider below, the data consist of 200 independent, identically distributed samples drawn from a data generating distribution P. Each observation consists of a single baseline variable V, a binary treatment A, and a binary outcome Y. First, we generate the baseline variables V from a mixture of two normal distributions with variance 1 and with probability 1/2 of being centered at 0 or 1. The randomized treatment A is then generated, with probability 1/2 of being 0 or 1, independent of V. Lastly, the outcome Y is generated according to a logistic regression model for the probability that $Y = 1$ given treatment A and baseline variable V. We consider three different such logistic regression models, each of which leads to a different data generating distribution. The first contains just the treatment: $P(Y = 1|A, V) = \text{logit}^{-1}(A)$; the second contains the treatment and the baseline variable as main terms $P(Y = 1|A, V) = \text{logit}^{-1}(A + V)$;

and the third contains both main terms and an interaction term $P(Y = 1|A, V) = \text{logit}^{-1}(A + V - AV)$. This completes the description of the three data generating distributions used below, which we call data generating distributions 1, 2, and 3, respectively.

We next define the three working models used by the above methods in our simulations. Methods M0, M4, and M5 require working models for the probability that the outcome Y equals 1, given the treatment A and the baseline variable V. We define Working Model 1 as

$$\text{logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV\right). \tag{7}$$

This is a correctly specified model for the data generating distributions above. Working Models 2 and 3 are misspecified models. Working Model 2 has the wrong functional form, and Working Model 3 uses a "noisy" version of the baseline variable V (to represent the effect of measurement error). These working models are fully described in Web Appendix A. We point out that methods M1 and M2 do not use working models. Also, the permutation-based method M3 requires working models for the probability that the outcome Y equals 1, given just the baseline variable V; we define working models for method M3, analogous to those defined above, in Web Appendix A, where we also give the R code for our simulations.

Table 1 gives the approximate power, based on 100,000 Monte Carlo simulations, of methods M0-M5, under the three data generating distributions and three working models defined above. The columns correspond to data generating distributions 1, 2, and 3, respectively. The rows are in blocks corresponding to Working Models 1, 2, and 3, respectively. Starting in the leftmost column, we see that all methods have roughly the same power (about 90%), except method M0 (regression-based method of this paper) and sometimes M3 (permutation-based) have less power (as low as 81%). As one goes from left to right in the table, corresponding to going from data generating distribution 1 to 2 to 3, all methods except M0 consistently lose power. This is because both the risk difference and odds ratio decrease as we go from data generating distribution 1 to 2 to 3, and all methods except M0 generally have less power at alternatives for which the risk difference and odds ratio are smaller. Method M0, in contrast, has more power than the other methods at data generating distribution 3 (representing interaction between the treatment and baseline variable). Method M0 is better able, at least in these simulation examples and the simulations in Web Appendix A, to take advantage of the interaction effect, since method M0 is based on estimated regression coefficients corresponding to both the main term A and the interaction term AV.

Comparing the power of methods M0-M5 when using correctly specified Working Model 1 vs. misspecified Working Models 2 and 3, we see that misspecification reduces the power of all methods (except the intention-to-treat method M1, which completely ignores baseline variables). See Web Appendix A for more simulations, in which the power of methods M0-M5 is compared for different sample sizes, data generating distributions, and working models. Also in Web Appendix A, we consider the following: Type I error at various sample sizes, the impact of using different sets of coefficients in hypothesis test (*), and the possibility of combining several test statistics based on different methods or working models.

7 Data Example from Randomized Trial

We apply the regression-based hypothesis test (*) to the data from a recently completed randomized trial. The Methods for Improving Reproductive Health in Africa (MIRA) randomized trial (Padian et al., 2007) investigated the effect of diaphragm and lubricant gel use in reducing HIV infection rates among susceptible women. 5,045 women were randomly assigned either to the active treatment arm (which we call the "diaphragm arm") or to the control arm. By the end of the trial, there were 158 HIV infections in the diaphragm arm and 151 HIV infections in the control arm. The intention-to-treat analysis for the risk difference yielded the 95% confidence interval (-0.02, 0.01). This is strong evidence that the intervention has little or no effect on overall HIV rates, but the question remained as to whether the diaphragm intervention may have an effect for some high-risk subpopulations. The principal investigators identified five baseline variables they thought indicative of HIV risk: age, condom use reported at baseline, prevalence of HSV[5] infection, a composite indicator of subject risk behavior, and a composite indicator of partner risk[6]. We denote these variables by V_1, V_2, V_3, V_4, V_5, respectively. Traditional subgroup analyses, which would test for an effect within subgroups defined by each such variable, would require adjustment for multiple testing. A single hypothesis test can have more power. We show next how to use the regression-based method of this paper to carry out such a single test. We note that since we are testing this hypothesis post-hoc (after having seen the data), we must interpret any results with much caution; the goal here is to illustrate the application of the regression-based method of this paper in a real data example. In general, one should prespecify such an analysis as a secondary analysis in the study protocol.

We now describe how a hypothesis test of type (*) from Section 4 can be

[5] herpes simplex virus 2

[6] The indicators of subject risk behavior and partner risk are defined in Table 3 on page 7 of (Padian et al., 2007).

Table 1: Power of methods M0-M5. Sample size is 200 subjects. The data generating distributions corresponding to each column and the working models used are described in Section 6. "C-M-H test" below is an abbreviation for "Cochran-Mantel-Haenszel test." Distributions 1, 2, and 3 are defined in the text.

	Power When Data Generated from:		
	Distribution 1	Distribution 2	Distribution 3
Hypothesis Testing Methods			
		Using Working Model 1 (*Correctly* Specified)	
M0: Regression Based	0.86	0.71	0.93
M1: Intention-to-Treat	0.93	0.76	0.52
M2: C-M-H Test	0.90	0.79	0.49
M3: Permutation Based	0.92	0.79	0.64
M4: Targeted MLE	0.92	0.83	0.54
M5: Aug. Estimating Fn.	0.92	0.83	0.53
		Using Working Model 2 (*Wrong Functional Form*)	
M0: Regression Based	0.85	0.65	0.60
M1: Intention-to-Treat	0.93	0.76	0.52
M2: C-M-H Test	0.91	0.73	0.48
M3: Permutation Based	0.81	0.67	0.48
M4: Targeted MLE	0.93	0.76	0.52
M5: Aug. Estimating Fn.	0.92	0.76	0.52
		Using Working Model 3 (*Containing Measurement Error*)	
M0: Regression Based	0.85	0.65	0.62
M1: Intention-to-Treat	0.93	0.76	0.52
M2: C-M-H Test	0.90	0.74	0.48
M3: Permutation Based	0.81	0.69	0.52
M4: Targeted MLE	0.92	0.78	0.52
M5: Aug. Estimating Fn.	0.92	0.78	0.52

applied to test the null hypothesis of no mean treatment effect within strata of the above baseline indicators of HIV risk. To carry out hypothesis test (*), we first need to specify a regression model for the probability of HIV infection by the end of the trial, given treatment arm A and baseline variables V_1, V_2, V_3, V_4, V_5. We used the logistic regression model

$$\begin{aligned} m(A, V|\beta) = \text{logit}^{-1}(&\beta_0 + \beta_1 A + \beta_2 V_1 + \beta_3 V_2 + \beta_4 V_3 + \beta_5 V_4 + \beta_6 V_5 \\ &+\beta_7 AV_1 + \beta_8 AV_2 + \beta_9 AV_3 + \beta_{10} AV_4 + \beta_{11} AV_5). \end{aligned} \tag{8}$$

We also need to pre-specify a set of coefficients corresponding to terms containing the treatment variable, to use in the test. We used all such coefficients (i.e. $\beta_1, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}$), but in general the question of which set of coefficients to choose in order to have the most power is an open problem and area for future research that we discuss in Web Appendix A. This logistic model was fit using standard software (see R code in Web Appendix E), yielding coefficient estimates and robust standard errors as given in Table 7 in Web Appendix E. Based on these estimated values, we constructed a Wald statistic, as described in Web Appendix B. This resulted in p-value 0.97. We note that the p-value is quite similar when using link functions other than the logit link; for the probit link, the p-value is 0.98, and for the complementary log-log link, the p-value is 0.97. The p-value when using a logistic regression model containing only main terms is 0.67.

The hypothesis test just described has asymptotically correct Type I error, even if the logistic regression model (8) is misspecified, under the assumptions given in Sections 3 and 4. Had this test rejected the null hypothesis, this would have been evidence that within some stratum of the variables V_1, V_2, V_3, V_4, V_5 (which are thought to indicate HIV risk), the diaphragm intervention has an effect. This information could be useful in assessing whether the diaphragm and gel provide protection against HIV in at least some circumstances, which would be an important result (for example, suggesting the potential role of the cervix, which a latex diaphragm is designed to protect, in HIV transmission).

We now consider applying the alternative methods M1-M5 given in Section 6 to this data example. Given that the intention-to-treat analysis effectively rules out there being a strong overall mean effect of the intervention, it does not make sense to apply methods M2, M4, or M5; this follows since these methods have low power at alternatives for which the overall risk difference is close to 0 and the overall odds ratio is close to 1, which is the case in this data set. It does make sense to apply method M3 (the permutation-based test of Rosenbaum (2002)), since this method can have adequate power at alternatives where there is no overall mean effect, as long as there is a strong effect within some stratum of baseline variables. We applied method M3 (see Web Appendix E for R code), resulting in p-value 0.88.

8 Discussion

Regression models are often used to analyze randomized trials. However, in medical studies, due to the complexity of relationships between variables, simple regression models may fail to accurately represent the true relationships between these variables. Thus, it is important to know whether hypothesis tests based on regression models will be robust to misspecification of these models. Our contribution in this paper is showing that for many commonly used regression models, hypothesis tests based on these models are guaranteed to have asymptotically correct Type I error, even when the model is incorrect, in certain types of randomized trials. We showed in Section 5 how to augment linear models so that they will have this robustness property. Our results provide a strong motivation for using robust variance estimates whenever model-based hypothesis tests are used.

A limitation of our results is that they are asymptotic, guaranteeing correct Type I error only in the limit as sample size goes to infinity. We examine Type I error of the methods considered in this paper (M0-M5 from Section 6), for sample sizes ranging from 200 to 400 subjects, in Web Appendix A; for all the methods, and for all the data generating distributions and working models we considered, the Type I error at nominal level $\alpha = 0.05$ was always at most 0.06, and most often was at most 0.05.

Another limitation of our robustness results is that they are only proved for data obtained from randomized trials in which, in addition to treatment being randomly assigned to study subjects, the data on each subject is (approximately) i.i.d. from an unknown distribution. This is a limitation since the subjects actually recruited in medical randomized trials are screened for entry criteria, and thus in many cases may best be considered a convenience sample (Freedman et al., 2007, appendix to chap. 27). We conjecture that our results will hold in such a setting, under the weaker set of assumptions made in (Freedman, 2008b), but this is an open problem. We note, however, that our assumption that subject data is approximately i.i.d. may not be as restrictive as it seems. For example, subject data may be approximately i.i.d. not from a distribution corresponding to the general population, but from a distribution corresponding to those in a population who meet the entry criteria of a trial and who would be willing to participate.

An alternative set of methods that adjust for baseline variables in randomized experiments and that use regression models as working models, are the exact permutation tests of Rosenbaum (2002). These methods have been shown to have correct Type I error, even when working models are misspecified, but under a different framework (called randomization inference) than used in this paper. We give a detailed comparison of the assumptions of this paper and the randomization inference assumptions of Rosenbaum (2002) in Web Appendix F. Because of this difference in assumptions, it is in general difficult to devise a fair comparison of our methods and the randomization inference methods of Rosenbaum (2002). However, in the special case in which the outcome is binary, the assumptions of this paper simplify (as described near the end of Section 3) and a direct comparison becomes possible. We were therefore able to include a permutation test of Rosenbaum (2002) in our simulation study comparing the power of various methods in Section 6. In our simulation study, the permutation test (M3) sometimes had more power and sometimes had less power than the regression-based test (M0) of this paper, depending on the underlying data generating distribution. We note that except in the special case of binary outcomes, the permutation-based methods of Rosenbaum (2002) will not have asymptotically correct Type I error for testing the hypotheses considered in our paper in many situations. (See Web Appendix F for an example.) This is not surprising since these methods are designed for testing hypotheses under a different framework than considered here.

We caution that the results in this paper apply to testing the hypothesis of no mean treatment effect within strata of selected baseline variables, but our results do not apply to estimation of these treatment effects. The reason is that under the alternative hypothesis, misspecified models can lead to distorted effect estimates. Such distorted estimates, however, can be beneficial for power of hypothesis tests when the distortion makes effect estimates more extreme (Robinson and Jewell, 1991). It is an open question how to decide which models will provide the most power for specific alternative hypotheses, assuming models may be misspecified.

Acknowledgements

Michael Rosenblum was supported by a Ruth L. Kirschstein National Research Service Award (NRSA) under NIH/NIMH grant 5 T32 MH-19105-19. Mark van der Laan was supported by NIH grant R01 A1074345-01. We would like to thank the MIRA study team and investigators for allowing us to have access to the MIRA trial data. The MIRA trial was funded by the Bill and Melinda Gates Foundation grant number 21082. We would also like to thank Estie Hudes, Torsten Neilands, and Oliver Bembom for their useful feedback.

Supplementary Materials

The Web Appendices are at
`http://people.csail.mit.edu/mrosenblum/robustregression.pdf`.

References

Berk, R. (2004). *Regression Analysis: A Constructive Critique.* Sage Publications, Thousand Oaks.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.

Freedman, D. A. (1997). From association to causation via regression. *Advances in Applied Mathematics* **18,** 59–110.

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

Freedman, D. A. (2008a). On regression adjustments to experimental data. *Advances in Applied Mathematics* **40,** 180–193.

Freedman, D. A. (2008b). Randomization does not justify logistic regression. *Statistical Science* **23,** 237249.

Freedman, D. A., Pisani, R., and Purves, R. (2007). *Statistics. 4th Ed.* W.W. Norton and Company, Inc., New York.

Hernández, A. V., Steyerberg, E. W., and Habbema, J. D. F. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology* **57,** 454–460.

Huber, P. J. (1967). The behavior of the maximum likelihood estimator under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1,** 221–233.

Iacono, A. T., Johnson, B. A., Grgurich, W. F., Youssef, J. G., Corcoran, T. E., Seiler, D. A., and et al. (2006). A randomized trial of inhaled cyclosporine in lung-transplant recipients. *New England Journal of Medicine* **354,** 141–150.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22,** 719–748.

McCullagh, P. and Nelder, J. A. (1998). *Generalized Linear Models.* Chapman and Hall/CRC, Monographs on Statistics and Applied Probability 37, Boca Raton, Florida, 2nd edition.

Moore, K. L. and van der Laan, M. J. (2007). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 215. http://www.bepress.com/ucbbiostat/paper215* .

Padian, N., van der Straten, A., Ramjee, G., and et al. (2007). Diaphragm and lubricant gel for prevention of hiv acquisition in southern african women: a randomised controlled trial. *The Lancet* **370,** 251–261.

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. (with errata). *Mathematical Modelling* **7,** 1393–1512.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics* **23,** 2379–2412.

Robins, J. M. (2004). Optimal structural nested models of optimal sequential decisions. *Proceedings of the Second Seattle Symposium on Biostatistics. (Eds.) D. Y. Lin and P. Heagerty. Springer. New York.* pages 6–11.

Robinson, L. D. and Jewell, N. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* **59,** 227–240.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17,** 286–327.

Rubin, D. B. and van der Laan, M. J. (2007). Empirical efficiency maximization. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 220. http://www.bepress.com/ucbbiostat/paper220* .

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27,** 4658–4677.

van der Laan, M. and Robins, M. (2002). *Unified Methods for Censored Longitudinal Data and Causality.* Springer-Verlag, New York.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 213. http://www.bepress.com/ucbbiostat/paper213* .

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, New York.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64,** 707–715.