Two-stage Regression for Treatment Effect Estimation

by

Joshua Kane Errickson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2016

Doctoral Committee:

Associate Professor Ben B. Hansen, Chair
Professor Robert W. Keener
Professor Edward D. Rothman
Professor Brian P. Rowan

for Nora and the little cupcake

ACKNOWLEDGEMENTS

I would like to thank my Advisor, Ben Hansen, for all his assistance and support. I would also like to thank the entire Statistics department: the faculty, fellow students and all the support staff; as well as all others across the university whom I have interacted with over the years. Finally I would like to thank my wife for her patience.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

We consider two variations of two-stage regression used to fit models. Two-stage least squares has seen a lot of usage in statistics and econometrics in the context of instrumental variables. (Historically in e.g. Wright et al. [69] and Theil [65], more recently in e.g. Angrist and Imbens [5] and Imbens [40], and seeing applied use in e.g. Burgess [16], Auger, Farkas, Burchinal, Duncan, and Vandell [7], Asongu [6] and many others.) We consider more general contexts.

When studying whether an intervention or other treatment has a significant effect on the response, a researcher may be further interested in whether those higher at risk of a negative response see more benefit from the intervention. For example, in Giné et al. [29], the authors are studying biometric identification in rural Malawi, and studying whether its use in the credit system will increase the rate of loan repayment. Their results claim that the effect of biometric identification is the largest among those who are least likely to repay the loan in the absence of the intervention. We discuss this example further in Section 2.3.1.

The natural way of fitting such a model is a two-stage modification of the Peters-Belson method (see Section 2.2.2). In the first stage, a prediction of the response in the absence of treatment is obtained. The second stage uses this prediction as a modification to the treatment effect, allowing the discovery of both the overall treatment

effect as well as the additional effect due the predicted risk. Useful in randomized trials, or observational data which are similar to those produced from randomized trials, this method has the additional benefit of separating the relationship between the predictors and the response from the effect of treatment on the response. The method requires a strong first stage model fit, and if such a fit cannot be obtained, this method should not be applied. An alternative framework to consider this method would be in measurement error literature.[18] This enhanced Peters-Belson methodology is discussed in Chapters II and III.

Consider instead a situation where the response is binary and we are examining the effect of some intervention. If there is a treatment effect, we may be interested whether the magnitude of the treatment effect on the probability of seeing a particular response depends on the probability of seeing the response in the absence of treatment. If the dependence is there, then logistic regression can fit the model. However, if there is no dependence, linear regression may be preferred. Usually when a response is binary, logistic regression is preferred over linear for a variety of valid reasons[22], thus not allowing this model to be tested. By using a two-stage regression, we can model the relationships between the response and the predictors and response and the treatment effect separately, to enable a linear relationship between treatment and response. This methodology is discussed in Chapter IV.

CHAPTER II

Peters-Belson with Prognostic Heterogeneity in Treatment Effect

2.1 Introduction

When considering the effectiveness of a treatment or intervention, a goal of interest may be identifying those who would most benefit from the treatment or intervention, known as effect modification.[58] One version of effect modification, subgroup analysis, separates the population into subgroups and estimates treatment effects for each. To be optimal, this assumes that the researcher knows and has access to the "correct" sub-grouping variables.[49] Alternatively, an unstructured subgroup detection method will lead to an inflated Type I error if corrected for multiplicity.[45]

One method to address subgroup analysis which has seen usage lately[24, 29, 30] involves inverting the question of interest. Rather than looking for those who would most benefit from a treatment, we can instead ask whether an individual's predicted outcome in the absence of treatment is related to the strength of the treatment effect. In an ideal situation, it could be possible to show that those most at risk of a poor response benefit greatest from the treatment. For example, in a study of classroom performance, we might be able to claim that those students who are most at risk of failing (e.g. those with the poorest predicted grades in the absence of treatment)

would show the most benefit from some alternative instruction method. Surely a result along these lines would be beneficial to an overburdened state government looking to target a treatment, or a budget-strapped administrator looking to cut costs by treating as few individuals as possible.

The methodology being used to address whether those highest at risk are most benefitted by a treatment is a two-stage variation of the Peters-Belson method. In the first stage, the predicted response in the absence of treatment is modeled using only the control group. In the second stage, the sample is partitioned into quantiles based upon predicted response to control, and the treatment effect is estimated in each quantile, using the predicted response to control as an estimate for potential response of the treatment group to the control. Alternatively, the second stage can include an interaction between treatment and predicted response to control, representing the additional effect. This continuous interpretation admits an easier analysis by avoiding edge effects, and will be considered going forward.

When performing this analysis, there are two choices to calculating the standard error for the interaction term, whether to account for the additional variation (i.e. the measurement error) in the first stage. Not accounting for the additional variation assumes all variables in the second stage regression model are measured without error. Taking a hint from instrumental variables literature [68], we claim that it is necessary to account for the first stage variability to obtain proper coverage and Type I error rates.

To estimate the standard error in the second stage, we use a sandwich estimator based on the estimating equations literature [19, 63], which has been shown to account for the measurement error in the second stage plug-in values.[50] Although we find conventional Wald-type intervals not to maintain proper coverage, we find that coverage is much better maintained in an elaboration of the conventional procedure furnishing a confidence interval by explicitly inverting a family of hypothesis tests.

To standardize nomenclature, we will call the general methodology a Peters-Belson method with Prognostic Heterogeneity, or PBPH method. We will refer to the implementation in existing literature as the uncorrected PBPH method, and our variation where we correct the standard error in the second stage due to the measurement error from the first stage as the corrected PBPH method. The corrected PBPH method could be described as a generalized score procedure[14, 31, 59].

The structure of this chapter is as follows. We first review the literature surrounding the use of this method in Section 2.2. The method originates from work by Peters [53] and Belson [11]. Finally, we discuss two approaches in the existing literature to correct inference in this procedure; an in-sample computational approach [1] and an out-of-sample first stage estimate [35].

Following this, we describe in Section 2.3 more detail and present the results from Giné, Goldberg, and Yang [29] which we use as our motivating example. We will use this paper to show how the method is used in literature, identify the issues we see in the method, and ultimately offer our corrected approach.

After discussing the methodology that is needed in Section 2.4, we present the corrected PBPH method that offers a proper level α hypothesis test and confidence interval in Section 2.5. We show simulation results and re-examine the data from Giné et al. [29] in Section 2.6. Finally, in Section 2.7 we give concise advice on implementation of our method.

2.2 Background

A brief introduction into causal inference will benefit later understanding. Additionally, the historical development of the Peters-Belson method allows us to understand the current motivation to use the PBPH method. Finally, we look at other approaches to addressing the issues we raise.

2.2.1 Causal Framework

The notion of causal inference and potential responses has a long history, but its modern interpretation starts with Rubin [60]. At the heart of any causal problem is the desire to answer the question "What response would we observe if the treatment of interest were applied versus if the treatment of interest were not applied?"

Let Z_i be a binary indicator of the intervention which individual i received, convention being that $Z_i = 0$ for those receiving the control and $Z_i = 1$ for those receiving the treatment. (For notation in this document, we will often use c and t instead of 0 and 1 respectively in indices for clarity.) Letting Y_{iz} be the response that individual i experienced upon receiving z, we want to gain insight into $Y_{it} - Y_{ic}$ for each individual i, called the treatment effect. This would be the end of inquiry since, barring idealized experimental designs, we only observe $Y_i = Y_{it}Z_i + Y_{ic}(1 - Z_i)$. This is the Fundamental Problem of Causal Inference.[37] Y_{it} and Y_{ic} are known as potential responses; formally, Y_{iz} is the potential response of individual i had they received z.

There have been many methods proposed to bypass this issue, two of the most common being the average treatment effect (ATE) and the effect of treatment on the treated (ETT). Put simply, ATE is $\mathbb{E}(Y_t - Y_c)$, the population difference in mean response of the treatment group vs the control group. ETT is $\mathbb{E}(Y_t - Y_c|Z = 1)$, restricting that difference to the treatment group. Note that the ATE requires estimating both $\mathbb{E}(Y_t)$ and $\mathbb{E}(Y_c)$, whereas when we restrict attention to the treatment group in the ETT, we have that $\mathbb{E}(Y_t|Z = 1) = Y$ and only need to estimate $\mathbb{E}(Y_c|Z = 1)$. The Peters-Belson method described below upon which our we based our work is aiming more towards the ETT than the ATE.

Causal inference enjoys a rich and deep literature, but this background should be sufficient at the moment for the problems at hand. For further details, see for example Pearl, Glymour, and Jewell [52] or Imbens and Rubin [41].

2.2.2 Peters-Belson Method

A natural estimate of the ETT is $Y_t - \hat{Y}_c$ amongst the treatment group. If we have confidence in $\hat{Y}_c$ as an estimate for Y_c, then we have confidence in the estimate of the treatment effect. Peters and Belson introduced a technique, now known as the Peters-Belson method, to obtain $\hat{Y}_c$ in two independent papers. The goal of Peters [53] was to introduce an alternative to pair matching that didn't have data loss to the same degree.[1] To do so, he used the control group data to fit a predictive model for the response, and then used that model to predict the responses of the treatment group data (i.e. predict $Y_{ic}, \{i : Z_i = 1\}$ using a model fit upon observations $\{i : Z_i = 0\}$). His novel claim was that these predicted responses were the response that the treatment group would have had were the treatment to have no effect, an idea that no doubt informed Rubin's later work. This can be followed by a trivial (Peters even thought so, as "if one uses a calculating machine, it moves very rapidly." [53, pp. 609]) test of differences between the average predicted and average observed response such as ANOVA.

Belson [11] had a similar goal as Peters, but his work included a bit more rigor. Specifically, his concern was the correlation between treatment status, Z_i, and covariates, $X_i \in \mathbb{R}^p$. The paper spends considerable time finding "stable correlates" in the data, covariates correlated to the response but unaffected by treatment status. These stable correlates would offer a strong predictive model of the response, and it is reasonable to assume they are balanced between the control and treatment groups. Belson found these stable correlates manually, a task which is no longer necessary due to increased computational power and continued work in the area of covariate balance.

[1] This was prior to the introduction of full matching, which has improvements over pair matching, including addressing the data loss.

Cochran [20] examined the Peters-Belson method in to determine when to use it over using some pooling method from the entire sample. Let $\hat{\beta}_c \in \mathbb{R}^p$ be the estimate of the coefficients on X from a predictive model fit from only the control group and $\hat{\beta}_t \in \mathbb{R}^p$ correspondingly from only the treatment group. Let $\hat{\beta} \in \mathbb{R}^p$ be the estimate using the entire sample. (X should not include the treatment indicator Z.) Cochran concludes that the only time it is not recommended to use $\hat{\beta}_c$ over $\hat{\beta}$ to obtain $\hat{Y}_c$ is when $\hat{\beta}_c$ and $\hat{\beta}_t$ are statistically indistinguishable and $\overline{X}_c$ and $\overline{X}_t$ are substantively different. In all other settings, notably anytime $\hat{\beta}_c$ and $\hat{\beta}_t$ are statistically significantly different which likely covers most data sets, the Peters-Belson estimate is recommended.

Similar methods have been discussed in the economics literature and are known as Oaxaca-Blinder methods, see Oaxaca [51] or Blinder [13].

2.2.3 Existing Approaches

There have been two general approaches in the literature to adjusting the PBPH method to account for the variation introduced from the first stage. The first uses computational methods and the second uses an out-of-sample alternative data set. We review both methods here and explain their limitations.

2.2.3.1 Computational Approach

This approach was identified in Abadie et al. [1]. The authors acknowledge the issues with current methods, and convincingly demonstrate their problems. After discarding the treatment group data from their motivational data sets, they perform a simulation by splitting the control group data into a faux treatment and faux control group. In this setting, the true treatment effect is zero, and thus there should be no ability to make any claim regarding the benefit of the treatment on those with the lowest predicted response. However, after performing the analysis using terciles based

upon the predicted outcome, the results show that the "treatment" is beneficial to those at highest risk and harmful to those at the lowest risk.

The authors suggest using some variant of cross-validation to correctly obtain proper coverage. Averaging over many repetitions, the authors find that either the leave-one-out or sample splitting variations of cross-validation yield estimators which obtain proper coverage.

However, these approaches introduce a computational complexity to a problem where none previously exists. While in most moderate sample size settings, any modern computer will be able to easily handle this approach, for larger samples, as with any bootstrap-based method, the computation time can easily transform from a minor nuisance into a major hindrance.

2.2.3.2 Out-of-Sample Approach

Another source of papers using this method are found in the medical literature. Hayward et al. [35] show that this two-stage approach to subgroup analysis has higher power as compared to traditional single variable sub-grouping. However, the authors also recognize the potential for the issues we discuss in this chapter, and recommended an out-of-sample solution, requiring that the first stage modeling of the predicted response be based upon an external data set or historical information. This is similar in concept to the sample splitting method of Abadie et al. [1] in that independent stages corrects the coverage.

Of course, in most situations, no such other data set or historical data exists to enable the independent modeling of the stages. If data do exist, it will require the often strong assumption that both data sets are from the same population. This assumption is a tempting one for researchers to make were this solution their only possible course of action.

2.3 Motivation and Initial Results

Here we present one of the papers which first brought this issue to our attention, and show via simulations that the issues we identify do exist. We follow this by discussing an alternative framework which may appeal to some and examine an additional concern, namely that the PBPH method, in addition to having an inflated Type I error, is also biased.

2.3.1 Empirical Example

One of the papers which motivates this work is Giné et al. [29]. In it, the authors are studying microloans, which are very small loans typically given to impoverished individuals (for an example in action, see www.kiva.org). Banks in developing countries where infrastructures such as photographic ID cards or personal biometrics do not exist can have trouble tracking individuals. Borrowers in default can visit different banks or bank officers and give new names, receiving loans that would not be given to someone with such a history of default. Some countries have started implementing massive programs designed to track borrowers, using fingerprinting or iris scanning or some similar method.

Honest brokers in both sides in the microloans transaction should benefit from the addition of the identification. The bank benefits by lowering its default rate. Those borrowers who are not prone to defaulting can more easily build a positive credit history, leading to loans with more favorable terms.

However, the authors note that there is so far little empirical evidence of the benefit of such a system. The authors performed a randomized experiment in rural Malawi. Farmers received microloans from a bank (in the form of credit at a local agricultural supply station, not cash) at the beginning of a growing season, and repaid the loan after the season's harvest was sold. The banks had the basics of a credit history system, but it relied on bank officers personal knowledge of individuals - a

farmer could easily go to another officer or bank as described above. The response of interest we will focus on is fraction of loan repaid on time.

The authors randomized all farmers who applied at the beginning of the season into two groups, a control group and a treatment group.[2] After an explanation of the benefits and punishments inherent in a proper credit history system, the bank took fingerprints from those in the treatment group. Until this point we have described a designed experiment rather than an observational study, thus minimizing any concerns about treatment assignment and covariate or response bias. However, the randomization is performed amongst farmers who apply for loans. After the randomization and application of treatment, the loan officers decided whether to offer loans, and then farmers decided whether to accept the terms. Roughly 1/6th of the farmers (520 of 3,082) ended up accepting the loans and participating in both the prior and post surveys.

First the authors fit a model predicting the response amongst the control group only, similar to

$$Y = X\beta + \epsilon, \tag{2.1}$$

where X is a matrix of baseline covariates, including an intercept. Thus, $\hat{Y}_c = X\hat{\beta}$ is the predicted potential response to control, amongst the entire sample. Next, fit two separate second stage models on the entire data. First,

$$Y = Z\gamma + \hat{Y}_c\rho + (Z\hat{Y}_c)\tau + \epsilon, \tag{2.2}$$

where $Z\hat{Y}_c$ is the interaction between treatment and predicted response. Secondly, for interpretability, they split the sample into quintiles based upon the predicted response

[2]In the actual study, the unit of randomization was a "club," a collection of farmers, who apply collectively and share liability in exchange for favorable lending terms from the bank. This added complication does not affect our general discussion or results, and thus is disregarded here. See the following chapter, specifically Section 3.5.3, where we re-introduce the clubs.

and fit

$$Y = Z\gamma + \sum_{i=1}^{5} \left[D_i \rho_i + (ZD_i)\tau_i \right] + \epsilon. \tag{2.3}$$

Here D_i is an indicator of membership in quintile i. This has the same basic idea of examining the interaction between treatment effect and predicted response, but admits an easier interpretation.

The results for the terms which the authors attach causal interpretations to are included in Table 2.1.

		Coef (SE)	
(Eq. 2.2)	Fingerprint	0.719 (.108)	***
	Fingerprint : Predicted Repayment	-0.807 (.120)	***
(Eq. 2.3)	Fingerprint : Quintile 1	0.506 (.125)	***
	Fingerprint : Quintile 2	0.056 (.105)	
	Fingerprint : Quintile 3	-0.001 (.048)	
	Fingerprint : Quintile 4	-0.040 (.044)	
	Fingerprint : Quintile 5	-0.075 (.044)	*

Table 2.1: Coefficient estimates for the two models taken from Giné et al. [29]. In the second model, quintile 1 contains individuals with the lowest estimated repayment rate, and quintile 5 contains those with the highest estimated repayment rate. The stars follow R notation, such that one (*) and three stars (***) indicates significance at the 10% and 1% level respectively.

The negative interaction effect from (2.2) and the pattern of interaction effects from (2.3) show what the authors were hoping for, namely that it appears those who are predicted to have the worst response are those whom the treatment helps most.

This agrees nicely with the intuition that farmers who already repay their loans don't need the extra incentive/threats, and that those who don't repay their debt are now forced to do so in order to continue obtaining loans. This knowledge could be beneficial to policy decisions, in that it may be easier to get a treatment approved which is most effective on those at the highest risk.

2.3.2 Computational Evidence

We now show that by not considering the error associated with predicted repayment from the first stage, the standard errors in Table 2.1 are underestimated, increasing Type I error. We can empirically show the existence of the issue by performing the uncorrected PBPH method in a setting where we know the true treatment effect and interaction.

If we take the control group from Giné et al. [29] and randomly split into a faux treatment and faux control group, the true treatment effect in any subgroup is fixed to zero. The true control group sample size is 563, noted to alleviate any concerns about small sample issues.

When we now perform the uncorrected PBPH method on this faux treatment and faux control groups, we know that the treatment effect in any sample subgroup should be, on average, zero. We focus here, and beyond, on the version of the second stage defined within Giné et al. [29] by (2.2), with a continuous interaction.

Performing the randomization into faux treatment and faux control groups 1,000 times, we reject the null hypothesis in 69.5% of the runs, much higher than the expected 5%.

2.3.3 An Alternative Framework

If the appeal to a causal framework does not convince the reader, we can reframe the PBPH method in terms of controlling for nuisance variables in a regression model. Consider a setting with response y and two independent variables, x_1 and x_2, where the former is the variable of interest and the latter is a nuisance parameter (e.g., x_1 is a treatment variable and x_2 is some demographic variable; though in this framework we need not assume that x_1 is categorical). Assume without loss of generality that y,

x_1 and x_2 are centered. The traditional least squares model for this setup would be

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{2.4}$$

where we simultaneously consider both controlling the nuisance parameter and estimating the effect of the variable of interest.

However, it may be beneficial to consider these two issues separately - firstly, removing from y the variance associated with x_2, and then following up by independently investigating the effect of x_1. To be more precise, in a first stage, we fit

$$y_i = \beta_2 x_{2i} + \epsilon_{1i}, \tag{2.5}$$

to obtain $\hat{\beta}_2$, and then as a second stage, fit

$$y_i - \hat{\beta}_2 x_{2i} = \beta_1 x_{1i} + \epsilon_{2i}. \tag{2.6}$$

The main benefit of this modular approach is that it allows us to perform model fit diagnostics on the first stage, and to gain confidence in our modeling of the nuisance parameters, before we approach analysis on the parameter of interest. We then have two models which each perform their sole job to the best of their ability, rather than a single model which attempts to satisfy two masters.

Cochran [20] considered the Peters-Belson method within this framework, and as mentioned above in Section 2.2.2, showed that it is preferable to utilize only the control group in (2.5) to obtain $\hat{\beta}_2$. To re-summarize Cochran's results, his claim is, except in cases where the $\hat{\beta}_2$ obtained from only the control group does not differ from the $\hat{\beta}_2$ obtained only from the treatment group and where the sample means of x_2 in the control group and treatment group are different, then using the $\hat{\beta}_2$ from the control group only is optimal.

The benefit of this view of the method is that it directly shows that the standard error calculations which consider only the variance in the second stage are incorrect. In the basic least squares model in (2.4), the closed form solution for the standard error of $\hat{\beta}_1$ is (with all the typical ordinary least squares assumptions)

$$\text{s.e.}(\hat{\beta}_1) = \sqrt{\frac{1}{n-3}\frac{\sum x_2^2 \sum r^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}}, \tag{2.7}$$

where r are the observed residuals. This standard error for $\hat{\beta}_1$ depends on x_2. However, if we disregard the measurement error on $\hat{\beta}_2$ while we look at the standard error for $\hat{\beta}_1$ from (2.6), then

$$\text{s.e.}(\hat{\beta}_1) = \sqrt{\frac{1}{n-2}\frac{\sum r^2}{\sum x_1^2}}, \tag{2.8}$$

and we lose this dependence (except through the residuals). This shows that the typical least squares regression standard error calculations will not suffice in a PBPH approach, thus further suggesting the need for a standard error calculation that will include the variance from both (2.5) and (2.6).

2.3.4 Relationship Between Bias and Model Fit

Many well-known estimates are biased, such as the traditional standard error estimate. However, the bias-variance trade-off often allows these biased estimates to be very practical. The estimate of the interaction coefficient in a PBPH approach is similar, in addition to having improper coverage, it is also biased.

The bias is not an issue in the final conclusion (see Appendix A.1). However, it can be educational to look at what settings yield larger bias. This bias does not affect all models equally. As might be suspected, models which correctly specify the set of independent variables minimize the bias. Unobserved variables or included noise variables increase the bias. Even in a correctly specified model, there is some level of bias.

First, consider when the first stage regression fit excludes informative unobserved variables. Formally we represent this as a decrease in model fit, measured by the R^2 statistic. The bias increases as the R^2 decreases.

The second effect is from including noise variables in the model. Regardless of the estimated coefficient on these variables (though perhaps not in the case where the estimated coefficient is identically zero), the bias increases with the introduction into the model of variables wholly unrelated to the response.

Of the two effects, the former, from unobserved informative variables, has a much larger effect than the latter, from included noise variables.

We ran a simulation designed to observe these effects. Using a sample size of 100, we generated 40 independent variables, X_1 through X_{40}, and then generated a response

$$Y_i = \sum_{j=1}^{20} \beta_j X_{ij} + \epsilon_i, \tag{2.9}$$

so that Y is a linear combination of the first 20 X_j but is independent of the remaining 20. ϵ_i in noise, drawn from $N(0, 1)$. Additionally, we generate treatment indicator z where $z_i = 0$ for the first half of the units and $z_i = 1$ for the remaining. Since $p(Y|z) = p(Y)$, the true treatment effect is 0.

We repeated this data generation 1,000 times. In each iteration, we performed the PBPH method 40 times, where method k includes in the first stage only $\{X_i : i \leq k\}$. Therefore, $k = 20$ is an oracle model which contains all informative variables and no noise, $k = 1$ is the least informative model, containing only a single informative variables, and $k = 40$ is the most over-saturated model, containing all informative variables but also all the uninformative ones as well.

Figure 2.1 shows the results. As you can see, as the model fit increases (i.e. as k approaches 20), the bias drastically drops to its minimum at $k = 20$. However, as the noise variables begin entering the model (i.e. as k increases above 20), the bias begins to increase, albeit slightly. And as mentioned before, even the perfect oracle

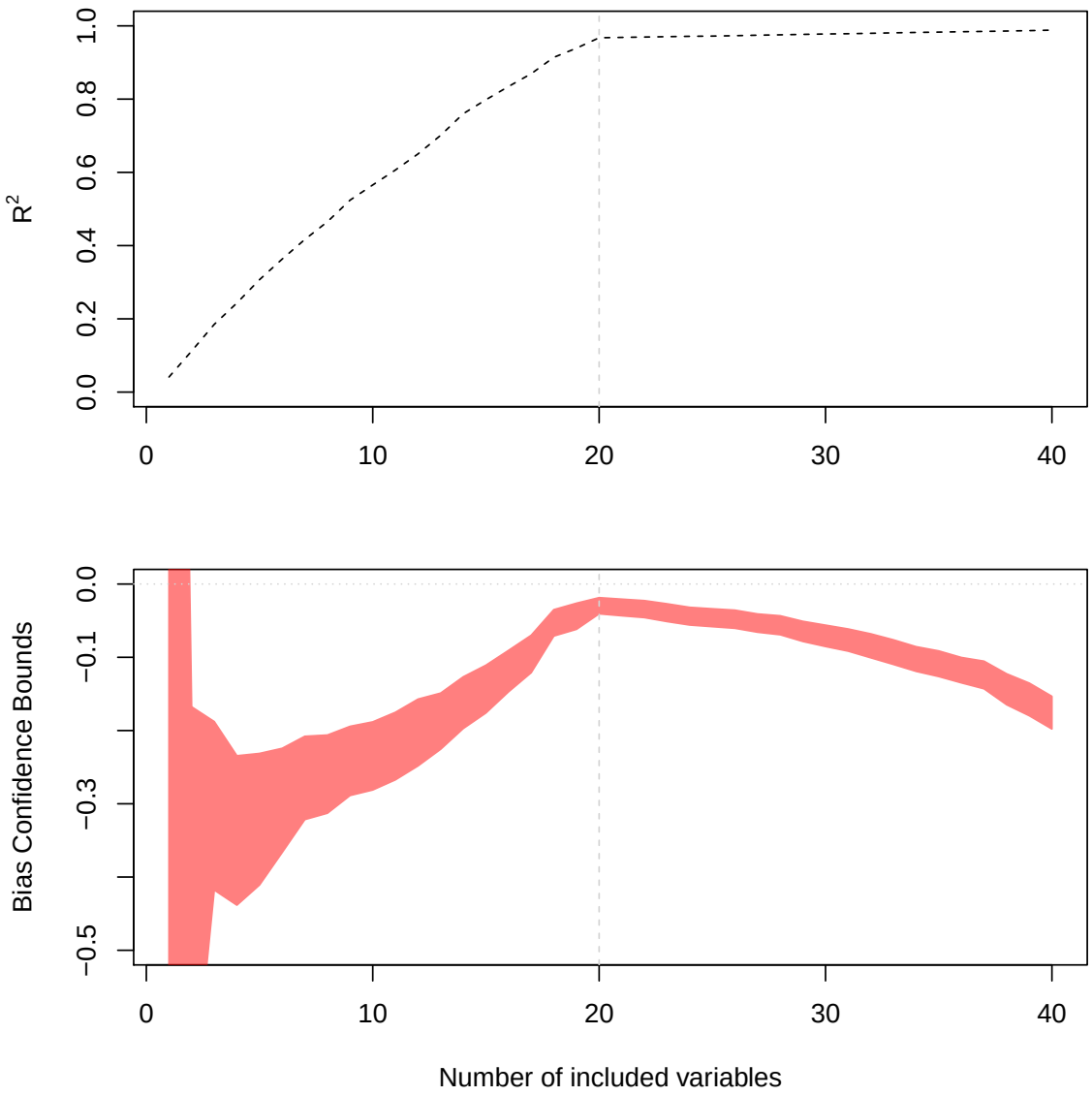

Figure 2.1: Visualizing the relationship between bias and the model fit. The true model includes only the first 20 variables, so the left side represents models with unobserved covariates, and the right side represents models with additional noise. Based upon 1,000 simulations at each number of variables.

model ($k = 20$) does not eliminate the bias entirely.

2.4 Methodology

Due to the issues we have identified, traditional methods for testing the coefficient on the interaction term in the PBPH method will not be sufficient. Specifically, the standard regression estimates of the variance of the coefficient will underestimate the truth, so we will turn instead to a robust sandwich estimator generated from estimating equations.

When we consider hypothesis testing, we will need to create an estimator for the covariance we have defined. There are several choices of such estimators, and we will describe the choices.

Finally, Wald-type confidence intervals will obtain incorrect coverage for the estimate of the interaction term. (See Appendix A.2 for details.) Generating a confidence region by test inversion instead allows us to obtain proper coverage.

2.4.1 M-estimators

M-estimators are a wide class of estimators which are useful in derivations of robust statistics. Each M-estimator is the solution to an estimating equation, namely $\hat{\theta}$ is an M-estimator for θ if $\hat{\theta}$ solves

$$0 = \sum_{i=1}^{n} \phi_i(X_i, \theta), \tag{2.10}$$

where X are some independent data and ϕ are known functions. Commonly, the right hand side is scaled by n to direct the conversation towards a mean and to ease some derivations, though consideration of sums helps our derivation.

We can place many common estimators into the M-estimation framework. For example, setting $\phi_i(X_i, \mu) = X_i - \mu$, it is easy to see that $\overline{X}$ solves $\sum_i \phi_i(X_i, \mu) = 0$,

and thus the sample mean is an M-estimator. The benefit of reframing estimators in this fashion is that it allows for more general asymptotic methods, as it can be shown under regularity conditions that M-estimators are asymptotically normal and consistent (when the error distribution is symmetric).[63]

We sketch a brief outline of the derivation of the asymptotic distribution of an M-estimator. There are many sources include the full derivation and proof such as Stefanski and Boos [63].

By Taylor expansion, the estimating equation (2.10) can be rewritten as

$$0 \approx \sum_i \phi_i(X_i, \theta) + \left(\sum_i \frac{\partial}{\partial \theta} \phi_i(X_i, \theta) \right) (\hat{\theta} - \theta). \tag{2.11}$$

In the limit, the remaining terms go to zero, provided certain conditions are satisfied. Stefanski and Boos [63] suggest a non-rigid version of these conditions: ϕ_i must be smooth and as $n \to \infty$, $\theta \not\to \infty$. For a more rigorous treatment of the conditions, see Huber [39] or Serfling [62].

Rearranging and in the limit,

$$\sqrt{n}(\hat{\theta} - \theta) = \underbrace{\left(\sum_i \frac{\partial}{\partial \theta} \phi_i(X_i, \theta) \right)}_{(*)}^{-1} \underbrace{\sqrt{n} \sum_i \phi_i(X_i, \theta)}_{(**)}. \tag{2.12}$$

When $\theta = \theta_0$, where θ_0 is the true population parameter, $(**)$ converges to normality with mean 0 and variance $\mathbb{E}\left[\phi(X_i, \theta_0)\phi(X_i, \theta_0)'\right]$. Call that variance the "meat," $M(\theta_0)$, which is the second non-central moment of the estimating equation, and is equivalent to the variance because the first moment is zero when $\theta = \theta_0$ by definition. Call $(*)$ the "bread," $B(\theta_0)$, which is the derivative of the estimating equation. Then, it follows from Slutsky's theorem that $\hat{\theta}$ is normal with expectation θ_0 and variance $B(\theta_0)^{-1} M(\theta_0) B(\theta_0)^{-T}$.[18, 63]

The bread is estimated by

$$B_n(\hat{\theta}) = n^{-1} \sum_i \frac{\partial}{\partial \theta} \phi_i(X_i, \hat{\theta}), \tag{2.13}$$

and the meat by

$$M_n(\hat{\theta}) = n^{-1} \sum_i \phi(X_i, \hat{\theta}) \phi(X_i, \hat{\theta})', \tag{2.14}$$

where $B_n(\hat{\theta})^{-1} M_n(\hat{\theta}) B_n(\hat{\theta})^{-T}$ converges in probability to $B(\theta_0)^{-1} M(\theta_0) B(\theta_0)^{-T}$ under regularity conditions.[42]

The sandwich estimator is a robust estimator, in the sense that consistency holds without any assumptions of distributions and even when the model is misspecified. However, when the model is correctly specified, the sandwich estimate is a very inefficient estimator.[19]

2.4.1.1 Stacked Estimating Equations

One limitation of classic sandwich estimators is the assumption that each ϕ_i has the same form. By using stacked estimating equation, we can bypass that limitation. This naturally arises in settings where an external data set estimates a parameter used in a model on another data set. In this case, it is not appropriate to discard the variation in the estimate of the parameter from the external data set.

To make the notion of stacked estimating equations concrete, let us assume that our model of interest has data X with parameter θ, and that parameter β comes from an external data set Y, so that our current model has dependencies on both θ and β, but the model on the external data only depends on β. In addition to $\phi_i(X_i, \theta, \beta)$, we can define an additional set of estimating equations, $\psi_j(Y_j, \beta)$. Then,

our M-estimators $(\hat{\theta}, \hat{\beta})$ are the solutions to

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \sum_j \psi_j(Y_j, \beta) \\ \sum_i \phi_i(X_i, \theta, \beta) \end{pmatrix}. \tag{2.15}$$

While the algebra becomes considerably more tedious, by setting up the bread and meat as blocked matrices, the derivation of the sandwich estimator is straightforward.

2.4.2 Estimating Covariance in Hypothesis Tests

The covariance matrix generated from the use of M-estimators is complex and careful consideration needs to be given to its estimation. Following the terminology and descriptions from Lindsay and Qu [46], we will mention three variations. Previously, in Section 2.4.1, we described (implicitly) two of these variations.

First, a model-based version of the covariance,

$$B(\theta_0)^{-1} M(\theta_0) B(\theta_0)^{-T}, \tag{2.16}$$

under the null hypothesis. This of course is only valid if the null hypothesis is correct, but minimizes additional sources of variation.[46]

Secondly, we have a fully empirical estimate, using (2.13) and (2.14), obtaining

$$B_n(\hat{\theta})^{-1} M_n(\hat{\theta}) B_n(\hat{\theta})^{-T}. \tag{2.17}$$

As mentioned above, with regularity conditions, we have that (2.16) converges to (2.17).[42]

Finally, we can use a hybrid of the model-based and empirical versions. In Lindsay and Qu [46], the variation used is a linear combination of the model-based and empirical estimators, e.g, if C_0 is the model-based version and $\hat{C}$ the empirical, a class

of hybrids $\hat{C}_0$ is defined as

$$\hat{C}_0 = (1-\alpha)C_0 + \alpha\hat{C}, \tag{2.18}$$

for $\alpha \in (0,1)$.

For sandwich estimators, an alternate form of the hybrid estimator can be defined in a simpler manner, by independently allowing the estimation of the bread and the meat by their respective model-based or empirical estimators. This leads to two alternative estimators,

$$B_n(\hat{\theta})^{-1}M(\theta_0)B_n(\hat{\theta})^{-T}, \tag{2.19}$$

$$B(\theta_0)^{-1}M_n(\hat{\theta})B(\theta_0)^{-T}. \tag{2.20}$$

We will seek guidance from simulations to compare the forms of the estimates.

2.4.3 Confidence Region by Test Inversion

The method is straightforward and based upon the duality of hypothesis testing and confidence intervals. If there is some test statistic $t(\theta)$ which at level α tests the hypothesis $H_0 : \theta = \theta_0$, rejecting when $t(\theta) > c^*$ where c^* is a critical value corresponding to the limiting distribution of $t(\theta_0)$, then a corresponding $(1-\alpha)\%$ confidence region for θ is

$$\{\theta : t(\theta) < c^*\}. \tag{2.21}$$

In general, the confidence region generated by test inversion need not be a continuous interval, but it often is.

Inverting a Wald test gives a Wald confidence interval. Let

$$t_W(\theta) = \frac{\hat{\theta} - \theta_0}{\sigma(\hat{\theta})}, \tag{2.22}$$

where $\sigma(\hat{\theta})$ is the sample standard deviation. Rejecting when $t_W(\theta) > z^*_\alpha$, we can

directly solve for θ_0, obtaining the traditional confidence interval of

$$\hat{\theta} \pm z_{\alpha}^{*} \sigma(\hat{\theta}). \tag{2.23}$$

However, consider a score test, with a test statistic of the form

$$t_S(\theta) = \frac{\hat{\theta_0} - \theta}{\sigma(\theta_0)}, \tag{2.24}$$

where the standard deviation is based upon θ_0, instead of $\hat{\theta}$. For tests of this form, it is not guaranteed that the test is invertible cleanly such that the confidence region will have a closed form solution. If such a closed form solution exists, it may not have the interpretability that (2.23) has. More generally, we can iterate over possible values of θ_0, and define the confidence region as all values of θ_0 for which (2.24) fails to reject.[3]

2.5 Calculations

Now that we have shown the issues in the PBPH method, and that the issues stem from an uncorrected standard error estimate, it remains to derive the corrected estimate. We will more rigorously define the problem before the derivation. Following that, we will examine how to perform hypothesis test and create confidence intervals.

2.5.1 Problem Definition

We first define the PBPH method rigorously.

Consider some data X of dimension $n \times p$ including a column of 1's for the intercept, and a response Y. Let Z indicate group membership; call $\{i : Z_i = 0\}$ the control group and likewise call $\{i : Z_i = 1\}$ the treatment group. Let $\sum_{i=1}^{n} Z_i = n_t$ and $\sum_{i=1}^{n} (1 - Z_i) = n_c$, with $n = n_c + n_t$.

In the first stage, we need a model fitted to predict the outcomes amongst only the control group. We derive this using a linear least squares model, but in principle, any model which can be used for prediction should suffice.

Within only the control group, fit

$$Y = X\beta_c + \delta, \tag{2.25}$$

where δ is the error term. The subscript c on the β_c coefficient is to remind that only the control group is used to generate it.

From the data, we obtain $\hat{\beta}_c$ as an estimator for β_c, and in turn, can obtain $\hat{Y}_{ic} = X_i'\hat{\beta}_c$, interpretable as an estimated potential response of observation i to the control. In the control group, $Y_i = Y_{ic}$, meaning the observed response is equivalent to the potential response to control, and thus $Y_i - \hat{Y}_{ic}$ is a residual. However, in the treatment group $Y_i = Y_{it}$ and therefore $Y_i - \hat{Y}_{ic}$ may be interpreted as an estimated treatment effect on individual i. The methodology suggested in Peters [53] and Belson [11] uses $n_t^{-1}\sum_{n_t}(Y_i - \hat{Y}_{ic}), \{i : Z_i = 1\}$ to estimate the treatment effect. That methodology assumes a homogeneous treatment effect.

To enable a heterogeneous treatment effect, introduce a second stage. To begin, we will utilize the full sample. The goal is to be able to make some statement speaking to the variation in the treatment effect with regards to the predicted response in the absence of any treatment. The right-hand side will be the observed response less the predicted response in the absence of treatment, $X\beta_c$. We will refer to this subtracted quantity as an offset. On the right-hand side, we have both the main treatment effect as well as the additional effect due to the predicted response in the absence of treatment. This model can be expressed as

$$Y - X\beta_c = Z\tau + (ZX\beta_c)\eta + \epsilon. \tag{2.26}$$

$Y = Y_t Z + Y_c(1 - Z)$ and thus

$$[Y_t Z + Y_c(1 - Z)] - X\beta_c = Z\tau + (ZX\beta_c)\eta + \epsilon, \tag{2.27}$$

is the true population model of interest. The left hand side is nothing more than the residuals left after (2.25).

When fitting this model on the entire data set, the control group will not affect estimates of τ or η since $Z_i = 0$ in the control group. We restrict attention to the treatment group, $Z_i = 1$, simplifying to

$$Y - X\beta_c = \tau + (X\beta_c)\eta + \epsilon. \tag{2.28}$$

To remove the dependence between τ and η, we center $X\beta_c$ on the right hand side of (2.28) relative to the treatment group to induce orthogonality. Let $\overline{(X\beta_c)_1}$ represent the mean of $X\beta_c$ amongst observations where $Z_i = 1$. Now (2.28) can be rewritten as

$$\begin{aligned} Y - X\beta_c &= \tau + \left(X\beta_c - \overline{(X\beta_c)_1}\right)\eta + \overline{(X\beta_c)_1}\eta + \epsilon \\ &= \left(\tau + \overline{(X\beta_c)_1}\eta\right) + \left(X\beta_c - \overline{(X\beta_c)_1}\right)\eta + \epsilon \\ &= \tau' + \left(X\beta_c - \overline{(X\beta_c)_1}\right)\eta + \epsilon. \end{aligned} \tag{2.29}$$

This gives us a more natural interpretation of the intercept. τ is the expected treatment effect when $X = 0$, which may not be an interesting value of X. However, τ' is the expected treatment effect when X is at its mean. The estimated treatment effect is equivalent to estimate from the methodology in Peters [53] and Belson [11], where $\hat{\tau}' = n_t^{-1} \sum_{n_t} (Y_i - X\hat{\beta}_c), \{i : Z_i = 1\}$. The estimate and interpretation of η does not change between (2.28) and (2.29).

To simplify notation forward, we will assume that $X\beta_c$ is centered as described

above such that $\tau = \tau'$, and that τ is therefore interpreted as the same treatment effect from Peters [53] and Belson [11].

2.5.1.1 Reasonable values for η

Since we will be fitting the second stage model only on the treatment group, (2.27) becomes

$$Y_t - X\beta_c = \tau + X\beta_c\eta + \epsilon, \tag{2.30}$$

in the population of treated individuals. When $\eta = 0$, the treatment effect τ is constant across all individuals.

Now, consider instead the case where $\eta = -1$. Then (2.30) becomes

$$Y_t = \tau + \epsilon, \tag{2.31}$$

implying that the response under treatment is constant across individuals, within individual error.

If $\eta < -1$, the relationship between Y_t and Y_c is inverted, so that the covariates X have directly opposite relationship on the response. For example, if age is positively associated with Y_c, for $\eta < -1$, age would be negatively associated with Y_t.

On the positive side, while we do not have a nice boundary condition as -1, large values of η are equally troublesome. Namely, for large values of η, the effect of the coefficients is magnified several-fold.

All three of these cases, while plausible, would represent a treatment effect outside the normal considerations, and outside the scope of this work. Therefore, we will limit our investigation to $\eta \in (-1, 2)$. The choice of an upper bound of 2 is somewhat arbitrary, but we feel represents a natural cut-off point for "large" positive values.

2.5.2 Standard Error Correction

2.5.2.1 Uncorrected Estimator

Before we derive the corrected standard error estimator, we can show the derivation of the uncorrected standard error estimator, both to provide a comparison point and to demonstrate a straightforward application of a sandwich estimator. Obviously this is not the only way (nor the simplest) to obtain this estimate, but as we show, it produces a more general estimate for the standard error from ordinary least squares that, with some assumptions, reduces to the form of the standard error calculated directly from the linear model.

This approach is to discard the variance introduced from the first stage model in (2.25) and focus solely on estimating the standard error of (τ, η) from (2.27). In this uncorrected approach, in the second stage model, $\hat{\beta}_c$ is considered fixed, so we start with a slightly modified version of (2.28),

$$Y - X\hat{\beta}_c = \tau + X\hat{\beta}_c\eta + e. \tag{2.32}$$

Consider both X_i and $\hat{\beta}_c$ as column vectors of height p, such that $X_i'\hat{\beta}_c$ is scalar. Following along with the typical derivation of estimating equations to solve linear regression, for example in Carroll et al. [18], we can define the estimating equation as

$$\psi_i(Y_i; \tau, \eta) = (Y_i - X_i'\hat{\beta}_c - \tau - \eta X_i'\hat{\beta}_c)\begin{pmatrix} 1 \\ X_i'\hat{\beta}_c \end{pmatrix}, \tag{2.33}$$

noting that $\psi_i \in \mathbb{R}_2$.

Therefore, our estimates of (τ, η) come from solving

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \sum_{\{i:Z_i=1\}} \psi_i(Y_i; \tau, \eta) = \Psi(Y; \tau, \eta). \tag{2.34}$$

The bread matrix is the expectation of partial derivative of the estimating equation, so

$$B^{(u)}(\tau,\eta) = B^{(u)}_{n_t}(\hat{\tau},\hat{\eta}) = \sum_{\{i:Z_i=1\}} \begin{bmatrix} 1 & X_i'\hat{\beta}_c \\ X_i'\hat{\beta}_c & (X_i'\hat{\beta}_c)^2 \end{bmatrix}. \tag{2.35}$$

The equality of $B^{(u)}(\tau,\eta)$ and $B^{(u)}_{n_t}(\hat{\tau},\hat{\eta})$ is immediately obvious since in the uncorrected derivation, we do not consider $\hat{\beta}_c$ as random. The superscript (u) is to indicate this is from the uncorrected approach.

Further, the meat matrix is the second non-central moment of Ψ, so

$$M^{(u)}(\tau,\eta) = \sum_{\{i:Z_i=1\}} \left(\mathbb{E}\left(Y_i - X_i'\hat{\beta}_c - \tau - \eta X_i'\hat{\beta}_c\right)^2 \begin{bmatrix} 1 & X_i'\hat{\beta}_c \\ X_i'\hat{\beta}_c & (X_i'\hat{\beta}_c)^2 \end{bmatrix} \right). \tag{2.36}$$

Since $Y_i - X_i'\hat{\beta}_c - \tau - \eta X_i'\hat{\beta}_c$ is an error term, it is centered, and thus we can estimate as

$$M^{(u)}_{n_t}(\hat{\tau},\hat{\eta}) = \sum_{\{i:Z_i=1\}} \left((Y_i - X_i'\hat{\beta}_c - \hat{\tau} - \hat{\eta} X_i'\hat{\beta}_c)^2 \begin{bmatrix} 1 & X_i'\hat{\beta}_c \\ X_i'\hat{\beta}_c & (X_i'\hat{\beta}_c)^2 \end{bmatrix} \right), \tag{2.37}$$

The estimate of the covariance matrix is thus

$$B^{(u)}_{n_t}(\hat{\tau},\hat{\eta})^{-1} M^{(u)}_{n_t}(\hat{\tau},\hat{\eta}) B^{(u)}_{n_t}(\hat{\tau},\hat{\eta})^{-T}. \tag{2.38}$$

If we make the assumption that errors ϵ_i are homoscedastic with common mean 0 and common variance σ^2 and that we have some $\hat{\sigma^2}$ as an unbiased estimator for σ^2, then since $Y_i - X_i'\hat{\beta}_c - \tau - \eta X_i'\hat{\beta}_c = \epsilon_i$, we have that its variance is σ^2. Then the meat matrix is nothing more than

$$M^{(u)}(\tau,\eta) = \sigma^2 B^{(u)}(\tau,\eta), \tag{2.39}$$

with a corresponding equality for the estimated version, so that (2.38) simplifies to

$$\hat{\sigma}^2 B_{n_t}^{(u)}(\hat{\tau}, \hat{\eta})^{-1}, \tag{2.40}$$

which is the covariance estimate derived directly from ordinary least squares.

2.5.2.2 Corrected Estimator

We can use stacked estimating equations to account for the additional variability introduced from the first stage model. We now have two different forms of the estimating equations,

$$\phi_i(Y_i; \beta_c) = (Y_i - X_i'\beta_c)X_i, \tag{2.41}$$

$$\psi_i(Y_i, \beta_c; \tau, \eta) = (Y_i - X_i'\beta_c - \tau - \eta X_i'\beta_c)\begin{pmatrix} 1 \\ X_i'\beta_c \end{pmatrix}, \tag{2.42}$$

where $\phi_i(Y_i; \beta_c) \in \mathbb{R}_p$ and $\psi_i(Y_i, \beta_c; \tau, \eta) \in \mathbb{R}_2$. ϕ represents the contribution to the variance from the first stage, while ψ represents the contribution from the second stage.

Therefore, estimators for the all parameters of interest, (β_c, τ, η), are solutions from

$$\begin{pmatrix} \mathbf{0} \end{pmatrix} = \begin{pmatrix} \sum_{\{i:Z_i=0\}} \phi_i(Y_i; \beta_c) \\ \sum_{\{i:Z_i=1\}} \psi_i(Y_i, \beta_c; \tau, \eta) \end{pmatrix} = \begin{pmatrix} \Phi(Y; \beta_c) \\ \Psi(Y, \beta_c; \tau, \eta) \end{pmatrix}. \tag{2.43}$$

Since there is a natural demarcation between the two forms of the estimating equations, we can approach this derivation in a blocked matrix format. The bread

matrix has the form

$$B(\beta_c, \tau, \eta) = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}\frac{\partial}{\partial \beta_c}\Phi(Y; \beta_c) & \mathbb{E}\frac{\partial}{\partial (\tau,\eta)}\Phi(Y; \beta_c) \\ \mathbb{E}\frac{\partial}{\partial \beta_c}\Psi(Y, \beta_c; \tau, \eta) & \mathbb{E}\frac{\partial}{\partial (\tau,\eta)}\Psi(Y, \beta_c; \tau, \eta) \end{bmatrix}, \quad (2.44)$$

where $B_{11} \in \mathbb{R}_{p\times p}$, $B_{12} \in \mathbb{R}_{p\times 2}$, $B_{21} \in \mathbb{R}_{2\times p}$ and $B_{22} \in \mathbb{R}_{2\times 2}$. To simplify notation going forward, the submatrices of the bread and meat are written succinctly. For example, B_{11} is shorthand for $B_{11}(\beta_c, \tau, \eta)$ and that $\hat{B}_{11}$ is shorthand for $B_{n_c,11}(\hat{\beta}_c, \hat{\tau}, \hat{\eta})$.

B_{11} is straightforward since it involves only the first stage, so

$$B_{11} = \hat{B}_{11} = \sum_{\{i:Z_i=0\}} X_i X_i'. \quad (2.45)$$

Since the first stage does not include (τ, η),

$$B_{12} = \hat{B}_{12} = 0. \quad (2.46)$$

B_{21} is slightly more complicated, since β_c exists in both stages,

$$B_{21} = \sum_{\{i:Z_i=1\}} \mathbb{E}\begin{pmatrix} -(1+\eta)X_i' \\ (Y_i - \tau - 2(1+\eta)X_i'\beta_c)X_i' \end{pmatrix}, \quad (2.47)$$

and is estimated by

$$\hat{B}_{21} = \sum_{\{i:Z_i=1\}} \begin{pmatrix} -(1+\hat{\eta})X_i' \\ (Y_i - \hat{\tau} - 2(1+\hat{\eta})X_i'\hat{\beta}_c)X_i' \end{pmatrix}. \quad (2.48)$$

Finally,

$$B_{22} = \sum_{\{i:Z_i=1\}} \mathbb{E}\begin{bmatrix} 1 & X_i'\beta_c \\ X_i'\beta_c & (X_i'\beta_c)^2 \end{bmatrix}, \quad (2.49)$$

and is estimated by

$$\hat{B}_{22} = \sum_{\{i:Z_i=1\}} \begin{bmatrix} 1 & X_i'\hat{\beta}_c \\ X_i'\hat{\beta}_c & (X_i'\hat{\beta}_c)^2 \end{bmatrix}, \tag{2.50}$$

The meat matrix $M(\beta_c, \tau, \eta)$ will be similarly blocked. The diagonal blocks, M_{11} and M_{22}, will be the variance of Φ and Ψ respectively. The off-diagonal blocks are unfortunately much more complicated, if for no other reason than issues of dimensionality. However, if we assume that the treatment and control samples are random samples drawn from an infinite population, the samples can be considered to be independent, implying a covariance of zero between them. Therefore $M_{12} = M_{21} = 0$.

M_{11}, being the variance of Φ, is

$$M_{11} = \sum_{\{i:Z_i=0\}} \text{Var}\left(Y_i - X_i'\beta_c\right) X_i X_i'. \tag{2.51}$$

Again, $Y_i - X_i'\beta_c$ is simply the error, which has a zero expectation, so we can estimate this with

$$\hat{M}_{11} = \sum_{\{i:Z_i=0\}} (Y_i - X_i'\hat{\beta}_c)^2 X_i X_i'. \tag{2.52}$$

The bottom right piece involves all three parameters of interest

$$M_{22} = \sum_{\{i:Z_i=1\}} \text{Var}\left((Y_i - X_i'\beta_c - \tau - \eta X_i'\beta_c) \begin{pmatrix} 1 \\ X_i'\beta_c \end{pmatrix} \right), \tag{2.53}$$

with corresponding estimate

$$\hat{M}_{22} = \sum_{\{i:Z_i=1\}} (Y_i - X_i'\hat{\beta}_c - \hat{\tau} - \hat{\eta} X_i'\hat{\beta}_c)^2 \begin{bmatrix} 1 & X_i'\hat{\beta}_c \\ X_i'\hat{\beta}_c & (X_i'\hat{\beta}_c)^2 \end{bmatrix}, \tag{2.54}$$

The covariance of (τ, η) is therefore the lower right 2×2 sub-matrix of

$$B(\beta_c, \tau, \eta)^{-1} M(\beta_c, \tau, \eta) B(\beta_c, \tau, \eta)^{-T}. \tag{2.55}$$

Rewriting each matrix in its blocked form, we can simplify to

$$\text{Var}\left(\tau, \eta\right) = B_{22}^{-1}\left(M_{22} + B_{21}B_{11}^{-1}M_{11}B_{11}^{-T}B_{21}^{T}\right)B_{22}^{-T}. \tag{2.56}$$

This derivation in based on a more general derivation in Carroll et al. [18, pp. 373]. That derivation has two additional terms, each of which includes B_{12}. In the specifics of our method, $B_{12} = 0$, so those terms vanish.

Since $B_{22}(\beta_c, \tau, \eta) = B^{(u)}(\tau, \eta)$ and $M_{22}(\beta_c, \tau, \eta) = M^{(u)}(\tau, \eta)$, this corrected variance is equivalent to the uncorrected variance plus an additional component relating to the first stage. This corresponds with intuition that the uncorrected standard error estimate is underestimating because it does not account for the measurement error of $\hat{\beta}_c$.[3]

The simplifying homoscedastic assumptions for this model are that errors d from the first stage have mean 0 and variance σ_1^2 and the errors e from the second stage have mean 0 and variance σ_2^2, and all are centered and each have appropriate estimators. Then,

$$\hat{M}_{11} = \hat{\sigma_1^2}\hat{B}_{11}, \tag{2.57}$$

$$\hat{M}_{22} = \hat{\sigma_2^2}\hat{B}_{22}, \tag{2.58}$$

and (2.56) simplifies to

$$\hat{\sigma_2}^2\hat{B}_{22}^{-1} + \hat{\sigma_1}^2\hat{B}_{22}^{-1}\hat{B}_{21}\hat{B}_{11}^{-1}\hat{B}_{21}^{T}\hat{B}_{22}^{-T}. \tag{2.59}$$

[3]On the topic of measurement error, we explored a variation of this method using regression calibration from the measurement error literature[18, Ch. 4] which would account for the measurement error on $X\hat{\beta}_c$ from the first stage model by way of a shrinkage factor. In simulation studies (similar to those described throughout Section 2.6) compared to the confidence intervals generated by the methodology we ultimately recommend, the confidence intervals generated by the regression calibration approach undercovered (providing only 80% coverage on average) and were 20% wider.

As in the unsimplified version, this is equivalent to the uncorrected variance in (2.40) with an additional linear term.

2.5.3 Hypothesis Testing

Before turning to a confidence interval generated by test inversion, we need to define hypothesis testing in this setting. We need nothing beyond the use of the corrected covariance calculations.

Define $H_0 : \eta = \eta_0$ for some $\eta_0 \in \Omega_\eta$ where Ω_η is the set of all possible values of η. We limit ourselves to $\Omega_\eta = (-1, 2)$ here. Let the set of unconstrained estimates of the parameters be $\hat{\lambda} = (\hat{\beta}_c, \hat{\tau}, \hat{\eta})$. Let the set of estimates of the parameters under the constraint imposed by H_0 be $\tilde{\lambda}_0 = (\hat{\beta}_c, \tilde{\tau}_0, \eta_0)$, where $\tilde{\tau}_0$ is the least squares estimate of τ under H_0. Estimates for β_c are not affected by constraints on η.

We need to consider the choice of which covariance estimate from Section 2.4.2. We will show below in Section 2.6.2.1 that the hybrid estimate in (2.20) is the simplest form which obtains proper coverage, so

$$\sigma^2_{\tilde{\lambda}_0}(\hat{\eta}) = B(\tilde{\lambda}_0)^{-1} M_n(\hat{\lambda}) B(\tilde{\lambda}_0)^{-T}. \tag{2.60}$$

Note that, considering the piece-wise definition of the bread and meat matrices in (2.45)-(2.53), η only enters into B_{21} and M_{22}. Therefore, $\sigma^2_{\tilde{\lambda}_0}(\hat{\eta})$ depends on η_0 only through the contributions from B_{21}.

We obtain the test that rejects H_0 if

$$\frac{|\hat{\eta} - \eta_0|}{\sigma_{\tilde{\lambda}_0}(\hat{\eta})} \geq z^*_{(1-\alpha/2)}. \tag{2.61}$$

2.5.4 Test Inversion

We can invert the test defined in (2.61). Using the hypothesis of interest defined above, $H_0 : \eta = \eta_0$, we can perform a search over the space of possible values of η_0 and the confidence region would be all η_0 for which we do not reject H_0. As in the hypothesis test, the version of the covariance estimate matters, and we will use the same as the hypothesis test, (2.60).

Beginning with (2.61) and rearranging,

$$|\hat{\eta} - \eta_0| \geq z^*_{(1-\alpha/2)}\sigma_{\tilde{\lambda}_0}(\hat{\eta}). \tag{2.62}$$

To ensure the resulting equation is nicely quadratic and to eliminate the troublesome L_1 norm, we square both sides to obtain

$$w_\alpha(\eta_0) := (\hat{\eta} - \eta_0)^2 - \left(\chi^2_{(1-\alpha)}(1)\right)^* \sigma^2_{\tilde{\lambda}_0}(\hat{\eta}) \geq 0. \tag{2.63}$$

Inverting the inequality, we obtain as a confidence region

$$r_\alpha(\eta_0) := (\eta_0 : w_\alpha(\eta_0) \leq 0)\,. \tag{2.64}$$

As mentioned, in general a confidence region generated by test inversion need not be a continuous interval. $w_\alpha(\eta_0)$ is quadratic in η_0. To see this, note that by using (2.60), η_0 enters the corrected standard error only through the bread, specifically linearly in B_{21}. Combining this with (2.47) and (2.56), we have that $\sigma^2_{\tilde{\lambda}_0}(\hat{\eta})$ is quadratic is η_0, implying $w_\alpha(\eta_0)$ is as well.

This leaves us with four potential shapes of confidence regions. Letting $c_1 < c_2$ be constant, we can have confidence regions of the form (c_1, c_2), $(-\infty, \infty)$, $(-\infty, c_1) \cup (c_2, \infty)$ or $(\varnothing)$. The finite continuous confidence interval is desired, and we will show during simulations that it is the most likely result.

An empty confidence set (rejecting η_0 for all possible values) is not possible. This should be clear considering (2.61), as when $\eta_0 = \hat{\eta}$, the left hand side is 0, which can never be rejected for any reasonable value of $\alpha < 1$.

The infinite confidence interval may appear daunting, but in practice has little difference than a very wide confidence interval. Two-stage least squares methods having "wide" confidence intervals is a known problem in the instrumental variables literature. When the instrumental variable is "weak" (a notion akin to the first stage model fit being poor), the standard errors in the second stage tend to be very large.[68] If the first stage model fit is poor, we do not have a model which can predict the response the treated group would have seen in the absence of any treatment. Given this, any claim to a traditional treatment effect is weak, and further claiming to be able to identify a secondary treatment effect would be even weaker. It is intuitive that in order to identify any information about an ETT effect, we must be able to estimate $Y_c|Z = 1$ well.

Disjointly infinite confidence regions appearing are an undesirable curiosity. However, such regions are infrequent in our simulation results (see Section 2.6.3).

Expressing $w_\alpha(\eta_0)$ in quadratic form is a non-trivial task.[4] However, since we know $w_\alpha(\eta_0)$ is quadratic, by solving $w_\alpha(\eta_0)$ for three values of η_0, we can fit a regression line with a quadratic term to obtain the numeric coefficients. This does not allow interpretation of the coefficients (to be able to firmly determine situations that cause each of the three shapes of confidence regions) but it does simplify computation by avoiding the need to iterate over all values of η_0.

[4]When deriving the coefficient on the quadratic term with the use of symbolic software, the resulting coefficient was half a page and interpretation was utterly hopeless.

2.6 Simulations

We first show that our approaches to hypothesis testing and confidence interval by test inversion produce proper coverage. We examine the existence of infinite confidence regions and suggest some guidelines for avoiding them. We then return to the Giné et al. [29] and compare uncorrected PBPH vs corrected PBPH results.

2.6.1 Data Generation

For all simulations in this section, we use a generalized data generation method which is described here. This methods allows us to specify parameters of interest, such as η, and also incidental parameters like τ and nuisance parameters like σ^2.

The covariates X are generated randomly from a Normal distribution, generally $N(0,1)$, and $X \in \mathbb{R}_{n\times q}$. n represents the total number of observations, and we will typically use $n = 100$ and $n = 1{,}000$ to represent a "small sample size" and "large sample size" respectively. q represents the number of covariates. Note that the q here describes merely the dimensions of the generated X, and it can (and often will be the case) that the response Y is generated by a data generating matrix of dimension $n \times p$, which is a subset of X, such that $p < q$. This distinction is why we use q to represent the dimension of X and p to represent the dimension of the data generating matrix.

Asymptotic theory for sandwich estimators allows q to grow along with n, establishing consistency results paralleling those of the classical fixed-p development under the assumption that $q^2 \log(q)$ is $o(n)$, i.e. $q^2 \log(q)$ is small in relation to n.[36, 54] We generate a finite sample rule of thumb in Section 2.6.2.2, which translates into choices of $q = 7$ and $q = 17$ for $n = 100$ and $n = 1{,}000$ respectively.

The treatment variable, $Z \in \{0, 1\}$, is randomly assigned with some probability p_Z of being assigned to treatment ($Z = 1$).

To generate responses, we need to specify four parameters, β_c, τ, η and σ^2.

β_c represent the coefficients on the data generating matrix used to generate the response. As before, typically p of these β_c will be non-zero, while the remaining $q-p$ are 0. We use this sparsity specifically to examine under-fitting and over-fitting, but more generally to avoid accusations of using only oracle models. Non-zero β_c's are randomly drawn from $N(0,1)$.

τ, representing the additive treatment effect, is in theory allowed to take any value, but in practice we restrict it. Due to X being $N(0,1)$, we chose τ on the same scale, either manually fixed or drawn randomly $N(0,1)$. In real-world data situations where τ is on a larger scale than X, this method may not be the best approach, as either there is a clean separation between Y_c and Y_t, in which case this method isn't necessary, or the data is extremely noisy, so that it's unlikely to be able to tease out the subtle effect the method is looking at.

η, our main parameter of interest, is restricted to $(-1,2)$ as described in Section 2.5.1.1. We either iterate over a grid on that range, to examine coverage as η varies, or we draw it randomly from $U(-1,2)$. We also force $\eta = 0$ occasionally to remove it from the model entirely.

Finally, σ^2, the variance of the error on the relationship between X and Y. We assumed standardization of X and Y such that σ^2 is 1.

With the parameters specified, we can generate the response. Akin to the method, we do this in two stages.

First, we generate Y_c in the entire population, using

$$Y_c = X\beta_c. \tag{2.65}$$

Recall again that $\beta_c \in \mathbb{R}_q$, but some subset of those β_c can be zero, so that Y_c is truly generated by $p \leq q$ subset of X.

Next, we generate Y_t,

$$Y_t = Y_c + \tau Z + \eta Z Y_c. \tag{2.66}$$

Finally, we set $Y = Y_{\text{obs}}$ by

$$Y_{\text{obs}} = Y_t Z + Y_c(1 - Z) + \epsilon, \tag{2.67}$$

where $\epsilon \sim N(0, \sigma^2)$. Z defines treatment status, specifically that $Z_i = 1$ implies membership in the treatment group and $Z_i = 0$ the control group.

If $\eta = 0$, then (2.66) is simplified to

$$Y_t = Y_c + \tau Z \tag{2.68}$$

so that there is only an additive treatment effect.

If τ is also set to 0, then $Y_t = Y_c$ and the treatment is completely ineffective.

We add the error to Y_{obs} instead of Y_c and Y_t to ensure homogeneity of the error; it should be easy to see that if we added the error to Y_c or both Y_c and Y_t, the error variance could differ drastically in the treatment and control groups.

Of course, Y_t and Y_c are discarded, and Y_{obs} is treated as the only observable response.

2.6.2 General Results

2.6.2.1 Choice of Covariance Estimate

In Section 2.5.3, we choose to use an estimator of η's standard deviation, (2.60), based upon a hybrid estimator (2.20). We justify that choice now. Recall that $\tilde{\lambda}_0 = (\hat{\beta}_c, \tilde{\tau}_0, \eta_0)$ is under the constraint $H_0 : \eta = \eta_0$ with $\tilde{\tau}_0$ being the least squares estimate of τ under that hypothesis, and that $\hat{\lambda} = (\hat{\beta}_c, \hat{\tau}, \hat{\eta})$ is unconstrained. We have the choice between using $\tilde{\lambda}_0$ over $\hat{\lambda}$ in both bread and meat (2.16); neither bread

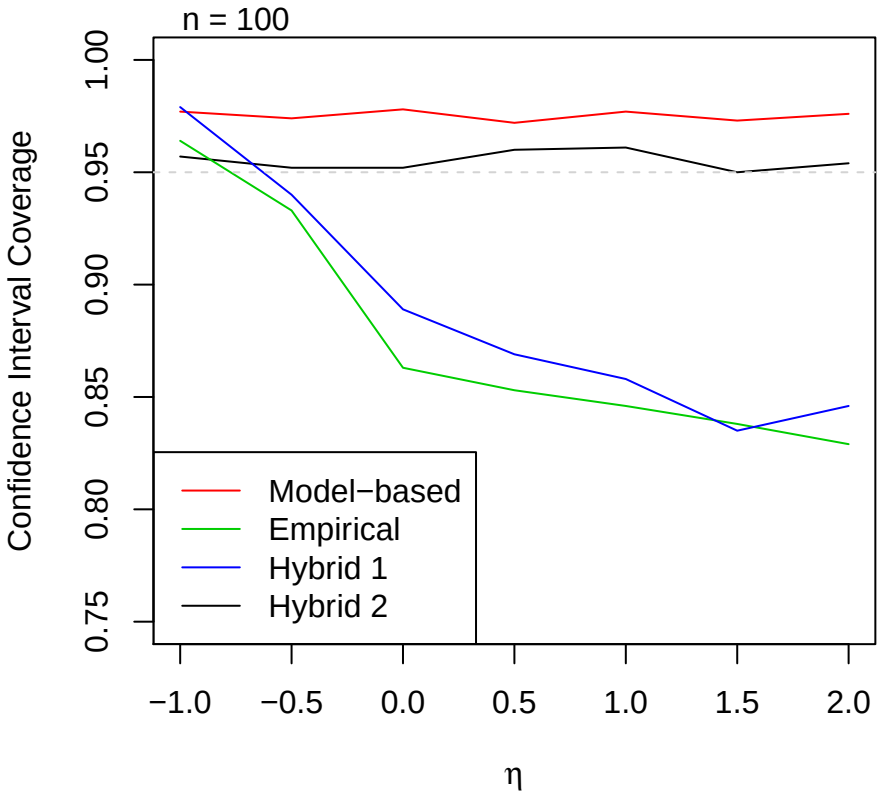

Figure 2.2: Choosing the version of the covariance estimator. Model-based refers to (2.16), empirical to (2.17) and the two hybrids are (2.19) and (2.20) respectively. Estimator Hybrid 2 obtains 95% proper coverage.

nor meat (2.17); or only meat (2.19) or only bread (2.20). We utilize the confidence interval instead of the hypothesis test to make this decision as the confidence interval by inversion contains, the hypothesis test of $H_0 : \eta = 0$. We perform simulations using each variation of estimate to examine coverage. The results are presented in Figure 2.2.

Neither the empirical version nor the first hybrid obtain proper coverage, which are the variations using $\hat{\lambda}$ in the bread. The model-based version and the second hybrid, which use $\tilde{\lambda}_0$ in the bread, obtain proper coverage, although the fully model-based version shows overcoverage. Using $\tilde{\lambda}_0$ in the bread adds stability to the covariance

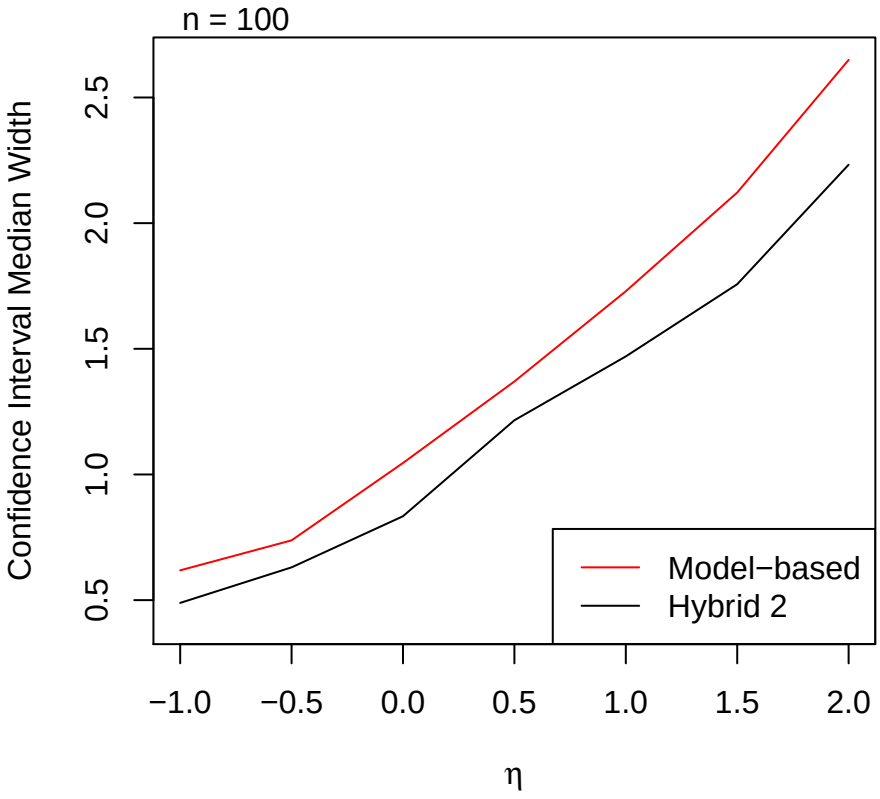

Figure 2.3: Median width of confidence intervals generated using model-based, (2.16), and the second hybrid, (2.20), which both obtained proper coverage. The model-based version averages 20% wider confidence intervals.

estimates, which is important because the bread is inverted.

The second hybrid is the best choice for three reasons. First, it obtains proper coverage without overcoverage. Second, the second hybrid estimator allows the covariance estimate to be quadratic in η_0, ensuring with simplicity that (2.64) defines a confidence interval. Finally, in lieu of power analysis we compare the two versions in terms of the width of the generated confidence interval. Figure 2.3 shows that the confidence intervals generated by the model-based version are 20% wider on average.

Therefore we have justified our choice in Section 2.5.3 of using

$$\sigma^2_{\tilde{\lambda}_0}(\hat{\eta}) = B(\tilde{\lambda}_0)^{-1} M_n(\hat{\lambda}) B(\tilde{\lambda}_0)^{-T}. \tag{2.69}$$

2.6.2.2 Finite Sample Rule of Thumb for q

We describe this as a rule of thumb to indicate that there is a certain amount of judgment behind the choice of the threshold; a more rigorous review of the topic may reveal a tighter rule. For our simulation results, this casual result is sufficient.

The asymptotic rule in He and Shao [36] is that $q^2 \log(q) = o(n)$. We consider a slower growth rate, $\frac{q^2 \log(q)^2}{n} < C$ and choose some C such that the rule holds. To determine the choice of C, we iterate over choices of n and C, and perform equivalence testing. Equivalence testing is used in clinical trials to test whether a new drug is not appreciably worse than an existing drug, as opposed to traditional superior hypothesis testing which considers whether the new drug is better than the existing.[67] Similar to that design, we wish to choose the largest C for which coverage is not significantly lower than $1 - \alpha$.

For our simulation, we run 10,000 repetitions of each (n, C) pair. We set a threshold of .1% as an equivalence region. This yields a rejection value of 94.7%.

The results of this simulation are displayed in Figure 2.4. The green squares have coverage above .947, and red below. Therefore, we choose $C = 2.5$, resulting in the aforementioned $q = 7$ and $q = 17$ for $n = 100$ and $n = 1{,}000$ respectively. It is likely there is another rule which is less strict as n increases, shown by the lack of failures for $n = 500$ or 1,000.

2.6.2.3 Hypothesis Test

We perform a level $\alpha = .95$ hypothesis test against $H_0 : \eta = 0$ by first generating data as described above, forcing $\eta = 0$, to ensure proper Type 1 error. The results for

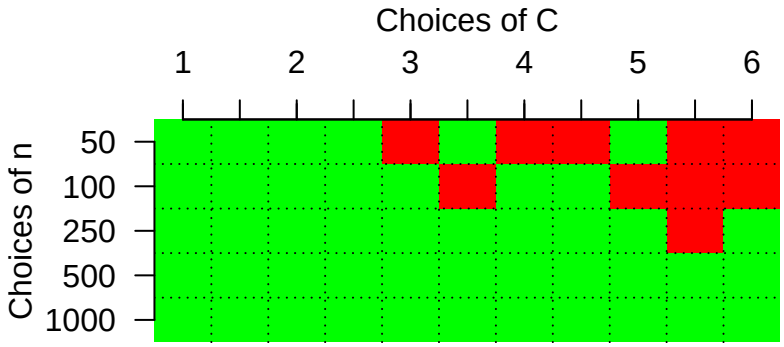

Figure 2.4: Simulation testing choices of C to establish the largest values of q. Each box represents 10,000 repetitions using the n and C combination. Red boxes reject the non-inferiority null that coverage with the (n, C) pair is no worse than 94.7%.

1,000 runs at $n = 100$ and $n = 1{,}000$ are in Table 2.2. The corrected PBPH method obtains proper coverage levels.

	n = 100	n = 1,000
Percentage Rejection	5.2%	4.7%

Table 2.2: Percentage of tests rejecting over 1,000 iterations.

2.6.2.4 Coverage Results

We've shown that the corrected PBPH method provides proper coverage on a level α test of the null that $\eta = 0$. The next step is to examine confidence interval coverage when η is not zero.

Table 2.3 shows coverage percentages across different values of true η. We obtain proper approximate 95% coverage across all reasonable values of η (again, see Section 2.5.1.1 for the rationale for the reasonable values) for both sample sizes. The negative bias discussed in Section 2.3.4 appears, though muted in the larger sample size.

Note that the overall coverage is for both desirable (continuous) and undesirable

(disjointly infinite) shapes of confidence regions combined. Table 2.4 shows coverage in each shape of the region. Amongst continuous confidence intervals, proper coverage is maintained. Amongst disjointly infinite confidence regions, we see over-coverage. Overall, in only a handful of the 7,000 total runs did we observe the situation where a disjointly infinite confidence region failed to cover the true value of η.

Table 2.4 does not include an entry corresponding to $n = 1{,}000$ because in the larger sample size setting, we did not encounter a single disjointly infinite confidence region in our simulations.

	$n = 100$		$n = 1{,}000$	
η	$\hat{\eta}$	Coverage	$\hat{\eta}$	Coverage
-1.0	-1.00	96.9%	-1.00	95.9%
-0.5	-0.55	94.8%	-0.50	95.1%
0.0	-0.10	95.6%	-0.01	94.2%
0.5	0.35	94.5%	0.49	95.3%
1.0	0.77	93.1%	0.98	94.1%
1.5	1.25	94.9%	1.48	95.1%
2.0	1.72	95.0%	1.97	94.5%

Table 2.3: Coverage of η, combined all shapes of confidence regions. For data generation, when $n = 100$ there are $q = 7$ parameters and $p = 3$, and for $n = 1{,}000$, $q = 17$ and $p = 6$.

	Continuous		Disjointly Infinite	
η	Count	Coverage	Count	Coverage
-1.0	992	97%	8	75%
-0.5	983	95%	17	100%
0.0	975	95%	25	100%
0.5	969	94%	31	100%
1.0	958	93%	42	100%
1.5	966	95%	34	100%
2.0	950	95%	50	100%

Table 2.4: Coverage of η over several values, separated by shape of confidence regions. $n = 100$ with $q = 7$ parameters and $p = 3$ used in data generation.

2.6.3 Infinite Confidence Regions

While in the previous section we considered both finite and infinite confidence intervals as similar, we separate them here to examine the relationship between first stage model fit and the shape of the confidence region. We restrict ourselves to smaller sample sizes as we never observed the disjointly infinite confidence regions in the n = 1,000 setting. We simulate data as described above with $n = 100$, but allow η to be drawn randomly from $U(-1, 2)$.

To measure model fit, we will consider the ANOVA associated with the first stage regression, and specifically its F-statistic. We base our choice on model fit measure due to guidance from the instrumental variables literature, particularly the determination of a weak instrument in the two-stage least squares instrumental variables procedure. For example, Stock and Yogo [64] derives a critical value to test directly against the first-stage F-statistic.

To ease interpretation, we'll look at the F-statistic on the log-scale. The results are presented in Figure 2.5. We can see that significance in the F-test is strongly associated with a finite confidence interval. Infinite confidence regions are almost entirely associated with a failure to reject the F-test. This corresponds to intuition, as if the first stage model fit is weak, we have little confidence in any second stage results.

Unfortunately, disjoint confidence regions are often associated with significant p-values, although not as significant as the finite confidence intervals. It would be convenient if the disjoint confidence regions were associated with poor first stage model fit, but since we are able to reject some values in the second stage, we must have some power at that stage. It follows that for the second stage to have power, the first stage must have had some power as well.

We looked at the additional measure of fit, R^2, and found an extremely similar pattern.

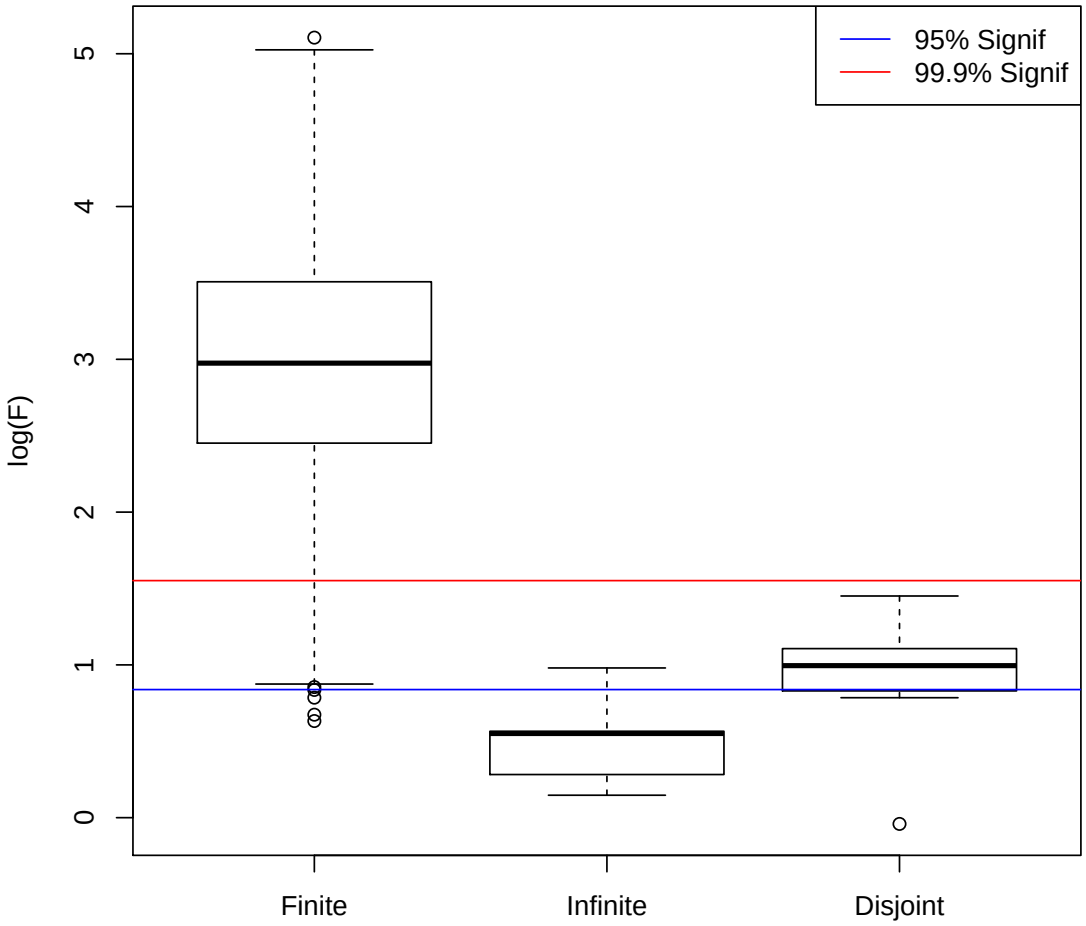

Figure 2.5: Comparison of F-statistic (on the log scale) across the different shapes of the confidence region.

Finally, and perhaps not surprisingly given the results thus far, if the first-stage model does not identify any non-zero predictors, obtaining a finite confidence interval is extremely unlikely.

2.6.3.1 An Attempt at Narrowing Infinite Confidence Intervals

As defined in Section 2.5.3, for each $H_0 : \eta = \eta_0$ which is tested, the parameters under the null hypothesis are $\tilde{\lambda}_0 = (\hat{\beta}_c, \tilde{\tau}_0, \eta_0)$, where $\tilde{\tau}_0$ is the least squares estimate of τ under H_0. Consider the model for obtaining $\tilde{\tau}_0$, fit only on the treated group,

$$\begin{aligned} Y - X\beta_c &= \tau_0 + \eta_0 X\beta_c + \epsilon. \\ Y - (1+\eta_0)X\beta_c &= \tau_0 + \epsilon. \end{aligned} \tag{2.70}$$

In other words, $\tilde{\tau}_0$ is the expected value of $Y - (1+\eta_0)X\beta_c$ among the treated. In the limits, we have that $\lim_{\eta_0 \to \infty} \tilde{\tau}_0 = -\infty$ and $\lim_{\eta_0 \to -\infty} \tilde{\tau}_0 = \infty$. If we consider τ as a nuisance parameter when testing hypotheses about η, then an approach introduced in Berger and Boos [12] suggests bounding $\tilde{\tau}_0$ by a wide confidence interval and appropriately adjusting the subsequent p-values. Assume that this approach would work, that is, that bounding $\tilde{\tau}_0$ by a wide (finite) confidence interval results in confidence intervals being finite. Because we are in theory (though not in practice) generating hypothesis tests over all values of η_0, and the asymptotic relationship between η_0 and $\tilde{\tau}_0$ is as described, bounding $\tilde{\tau}_0$ at any finite limits will equally result in finite confidence intervals for $\hat{\eta}$, and we need not restrict the bounds to the confidence interval of $\tilde{\tau}_0$.

Assume we bound $\tilde{\tau}_0 \in [u, l]$, that η_l and η_u solve $\tilde{\tau}_0 = l$ and $\tilde{\tau}_0 = u$ respectively, and that $\eta_u \leq \eta_l$ without loss of generality. With this modification, $w_\alpha(\eta_0)$ is no longer necessarily continuous, as it may be different in the three ranges $(-\infty, \eta_u)$, (η_u, η_l) and (η_l, ∞). However, the shape of $w_\alpha(\eta_0)$ is still quadratic in η_0 in each range.

The conjecture is that bounding $\tilde{\tau}_0$ would reduce the incidence of infinite confidence regions. If we rewrite $w_\alpha(\eta_0)$ as $a\eta_0^2 + b\eta_0 + c$ where a, b and c are some function of the data, the critical value of the χ^2 distribution, $\hat{\beta}_c$, $\tilde{\tau}_0$ and $\hat{\eta}$, then if $a > 0$, we must have a finite confidence region (because empty confidence regions are not possible, see the argument in the Section 2.5.4). However, we can easily find counterexamples. Table 2.5 shows a few. In each case, $n = 100$ and the data is generated as described in Section 2.6.1. Each simulation run resulted in an infinite confidence region and should be eligible to be adjusted by this approach. In each counterexample case, we see the coefficient on η_0^2 remain negative even as $\tilde{\tau}_0$ is bound.

	Left-asymptotic region	Center region; unbound $\tilde{\tau}$	Right-asymptotic region
1	-0.8η_0^2 - 7.2η_0 - 15.5	-0.4η_0^2 - 0.7η_0 - 0.4	-0.8η_0^2 + 2.7η_0 - 2.5
2	-0.4η_0^2 - 3.2η_0 - 5.5	-0.3η_0^2 - 0.9η_0 - 0.6	-0.4η_0^2 + 0.4η_0 - 1.1
3	-12.9η_0^2 - 117.8η_0 - 415.9	-8.7η_0^2 - 14.5η_0 - 8	-12.9η_0^2 + 49η_0 - 134.8

Table 2.5: Showing counterexamples to bounding $\tilde{\tau}_0$. Each equation is $w_\alpha(\eta_0)$, and the middle column represents both the unbounded version and the center region while bounding. The left- and the right-asymptotics are with the bounding. Bounding does not induce the coefficient on the squared term to be positive, and thus does not stop infinite confidence regions.

2.6.4 Underfitting and Overfitting in the First Stage

We've shown that fitting the first stage model well is important to being able to identify the additional treatment effect of the predicted response to control, if any. We now examine the effects of over-fitting on the first stage.

To start, we fix $n = 100$ and $q = 7$. However, when generating Y, we allow only the first four coefficients to be non-zero, hence $p = 4$. This allows us to compare three

difference models,

$$Y = \beta_1 X_1, \tag{2.71}$$

$$Y = \sum_{j=1}^{4} \beta_j X_j, \tag{2.72}$$

$$Y = \sum_{j=1}^{7} \beta_j X_j. \tag{2.73}$$

Clearly, the first model is underfit and the last model is overfit, while the middle model is an oracle model. We draw τ and η randomly, to disallow coverage being affected by the specific choices of τ and η. We repeated the above 1,000 times. In Table 2.6 we show overall coverage in each model and Table 2.7 shows the distribution of the shapes of the confidence region by model.

	Underfit	Oracle	Overfit
Coverage	97%	96%	95%

Table 2.6: Coverage of confidence regions based upon (2.71), (2.72) and (2.73) respectively, regardless of the shape of the confidence region.

	Underfit	Oracle	Overfit
Continuous	80%	98%	99%
Disjoint	20%	2%	1%

Table 2.7: Percentage of each type of confidence region found in simulation with each version of the model.

There are two notable conclusions to draw from this simulation.

First, underfit models in the first stage yield a slightly conservative coverage compared to oracle and overfit models. This is likely related to the extremely large percentage of infinite confidence intervals, which have no chance of rejecting the null hypothesis.

Secondly, within the restrictions on the dimension of q in Section 2.6.2.2, there is no penalty for overfitting, with equivalent coverage and very slightly less common

infinite confidence regions. When overfitting beyond the limits of those restrictions, we did see a lack of proper coverage on other simulations.

2.6.5 Returning to Giné et al. [29]

We return now to our motivating example, the microlending paper Giné et al. [29]. Recall that in this paper, the authors used an uncorrected PBPH method to estimate the effect of fingerprinting after stratifying the subjects into quintiles based upon their predicted loan repayment.

First, we repeat the simulation described in Section 2.3.2, where we split the control group into faux treatment and control groups. Then, knowing that all treatments effects are zero on average, we repeated the uncorrected PBPH method and rejected the null in 69.5% of the runs. In those same runs, utilizing the corrected PBPH method, we rejected the null in only 11.5% of all runs

Moving away from simulations, we compare the published results with the results from a corrected PBPH method.

As Table 2.1 cited, the reported coefficient for an interaction term between fingerprinting and a continuous predicted repayment was -0.807 with a standard error of 0.120. This standard error was the result of a bootstrap estimation procedure, and involved a model which included club effects (a club being a group of individual participants who assume joint risk for the loans in exchange for improved rates). Our results, contained in Table 2.8, directly estimate the standard error, and discards the club level effect for simplicity, so our results are slightly different than those published.

	Estimate	S.E.	Confidence Interval
Uncorrected	-0.896	0.043	(-0.980, -0.812)
Corrected	-0.896	0.054	(-0.998, -0.781)

Table 2.8: Comparison of uncorrected and corrected confidence intervals. Results differ from published results in Giné et al. [29] in Table 2.1 due to simplifying model.

This does not change the results of the paper; with the corrected PBPH we still

reject the null hypothesis $H_0 : \eta = 0$.

Notice that the standard error from the corrected method is larger and the confidence interval wider, and the corrected confidence interval is not centered around the point estimate. Theses are all behaviors predictable by our method; the first two due to considering the measurement error on the predicted response to control, and the last due to the creation of a non-Wald-style confidence interval.

2.7 Method Summary

Based upon the proceeding results, we summarize and recommend the following approach.

The first stage model should predict response amongst the control group only. The practitioner should strive to fit the first stage model as well as possible. Overfitting, up to the limits of the rate in Section 2.6.1, is not a concern. If good first stage model fit cannot be obtained, the conclusion should be that the data is inadequate to examine any additional treatment effect beyond the average treatment effect or effect of treatment on the treated.

Once a suitable first stage model is found, use its coefficients to predict the response in the absence of control amongst the treatment group members. This $\hat{Y}_c$ can be differenced from $Y = Y_t$, and regressed against $\hat{Y}_c$. After performing this regression, give proper consideration to the standard error and hypothesis test. Both should be computed as described earlier in the chapter.

The constant in the resultant second stage model can be interpreted as the main treatment effect. The coefficient on $\hat{Y}_c$ can be interpreted as an additional effect due to predicted response in the absence of treatment.

Due to the rarity of the disjointly infinite confidence region and the difficulty in interpretation of such a region, we recommend considering such a result as equivalent to a continuous infinite confidence interval. This will make coverage very slightly more

conservative (for example, in one run, coverage increased from 95.4% to 95.7%). We recommend not generating confidence intervals if the hypothesis test fails to reject, to help minimize the complications here.

If this recommendation is accepted, the confidence interval can always be considered continuous. If the confidence interval is wide, a next step should be to strengthen the first stage. Failing that, the conclusion should be that we can find no significant evidence of a interaction between treatment effect and predicted response in the absence of treatment.

We have implemented this suggested methodology in an R package **pbph**. All simulations were performed using the `pbph` function in this package.

2.8 Conclusion

We have introduced an analysis to answer a question which is popular in applied literature; are those more at risk benefited most by a treatment? After correcting the standard error calculations, we found that an ordinary Wald-style confidence interval was not sufficient. Our method considers multiple hypotheses about the parameter of interest by performing a test inversion, which does lead to a slightly more complicated approach.

On the other hand, following our generalized score procedure includes several advantages. First, by not having to resort to a profile likelihood style approach, we have only a single parameter of interest (η) to consider, instead of a parameter of interest which is dependent on the nuisance parameter (τ).

Secondly, and related to this, we are not required to fit multiple second stage regression models, saving substantial computational complexity.

Finally, because we have shown that our test statistic is quadratic (Section 2.6.2.1), we avoid an exhaustive search over the parameter space as is common in test inversion settings.

CHAPTER III

Further Applications of the Peters-Belson with Prognostic Heterogeneity Method

3.1 Introduction

In Chapter II, we introduced the Peters-Belson with Prognostic Heterogeneity method. The PBPH method addresses whether an intervention is most effective amongst those who are most likely to have a negative response in the absence of the intervention. The PBPH method is a two-stage modification of the Peters-Belson method. The first stage predicts the response in the absence of treatment using only control group members; the second stage models the treatment effect as the sum of a main effect and an additional effect due to the interaction of the treatment indicator and the predicted response from the first stage. As is common with two-stage regression procedures, the standard error associated with the estimated coefficients in the second stage has to account for the measurement error inherent in using a predicted response from the first stage, which we addressed with the use of a sandwich estimator based upon estimating equations. Following this, we showed the need to generate a confidence region via test inversion, as a Wald-style confidence interval produced undercoverage.

We introduce the **pbph** package implemented in *R*[55]. The package focuses on

implementing the second stage, allowing users freedom to create the first stage as desired. We extend the implementation of Sandwich estimators found in the **sandwich** package[70] to easily generate correct standard errors. Additionally, we efficiently implement the generation of the confidence region by test inversion, not requiring iterating over all possible values of the null hypothesis.

We allow two further complications to the method which the practicing statistician is likely to encounter.

First, in the previous chapter, the relationship between the response variable Y and its predictors was considered to be linear, and the error in the model assumed normal. This led to both stages being fit with linear regression. If Y were for example binary, we would prefer the first stage to be logistic. We extend our method to allow this modification by allowing the first stage to be fit with a generalized linear model. The **sandwich** package which we extend is generalized to many variations of model, simplifying this stage.[71]

Secondly, clustered random trials can be used in place of simple random assignment. For example, consider a population of students at a particular school; each class could be a cluster. Common clusters amongst larger populations include census tracts or congressional districts. In clustered random sampling, treatment is assigned at the cluster level instead of the individual level. However, this form of clustering introduces correlation amongst observations, as units within a cluster are typically more alike than units across clusters.[28] The traditional method of dealing with clustered standard errors is sandwich estimators.[17] We overload the `meat` and `sandwich` functions from the **sandwich** package to accept an argument identifying clusters.

3.2 Implementation

To call the main function of the **pbph** package, `pbph`, the first stage model must be fit by the user. This is fit using built-in *R* functionality, usually the `lm` function

(see Section 3.3.1 for the ability to fit the first stage via `glm`) and should only be fit on control group members. One of the benefits of the PBPH method is separating the goal of predicting the response in the absence of treatment from the goal of estimating any treatment effects. By enabling users to generate their first stage model fit externally, we allow them to create a model with a sole goal, rather than attempting to simultaneously also capture the treatment effect. Time and care should be taken at this step, as the stronger the first stage model fit, the more likely that the second stage model will be informative. See Section 2.6.3 which examines the existence of infinite confidence intervals in the presence of poor first stage model fit.

The function `pbph` takes in three arguments. The first is the first stage model fit as described. The additional arguments are a `data.frame` containing the data, and a treatment indicator (either a variable name inside the data or a `vector` of the same number of rows as the data) which assigns a `1` to each member of the treatment group and a `0` to each member of the control group. `pbph` follows other methodology which are elaborations on ordinary least squares, as implemented via `lm`. Some of these elaborations use `lm` explicitly by extending the `lm` class, such as `glm` from **stats** for generalized linear models [55] or `ols` from **rms** for saving design elements from a linear model [33]. Others could be called spiritual successors to `lm` as while they don't extend `lm`, their input and output function similar to `lm`, such as `lmer` from **lme4** for mixed models [10], `coxph` from **survival** for Cox regression [66] or `tsls` for two-stage instrumental variable regression from **sem** [26].

`pbph` itself is a very simple function. It generates the second stage model and saves it as an object of class `pbph`, which contains `lm` as described above. It returns the `pbph` object which contains a few additional pieces of information which are used in calculating standard errors, performing hypothesis tests and generating confidence intervals.

Further following along with `lm`, we do not generate the standard error yet. In-

stead,it is generated on demand, either when the user wishes to view it via `vcov` or when it is needed for calculations, for example in `summary` (which also performs a hypothesis test on each parameter) or `conf.int` to generate confidence intervals.

3.2.1 Standard Error Calculations and Hypothesis Testing

The standard error calculations utilize the existing `bread` and `meat` functionality from the **sandwich** package. Recall that we treat the bread and meat matrices as block matrices, each with four blocks. However, the off-diagonal blocks of the meat and the top right block of the bread were 0, leaving us with five pieces of bread and meat to calculate.

Both diagonal pieces of the bread are very straightforward, merely the matrix multiplication of the transpose of the design matrix by itself. In `R`, this is represented by

```
crossprod(x)
```

For B_{11}, the bread corresponding to the first stage model, `x` is X, the covariates, with the first column of $\mathbf{1}$. In B_{22}, the bread corresponding to the second stage, `x` is $(\mathbf{1}\ \mathbf{X}\beta_{\mathbf{c}})$, a column for intercepts and a column for the predicted response to control.

Similarly, the blocks of the meat are straightforward, requiring only the extra step of first generating the estimating function for the data, which is done using the `estfun` function of **sandwich**, prior to performing the matrix multiplication,

```
crossprod(estfun(x))
```

The remaining off-diagonal block of the bread requires more attention, both because its calculation is not as clean and because it has a dependency on η. As a result, the diagonal blocks need be computed only once, but the off-diagonal bread block varies with η_0. To calculate the hypothesis test of $H_0 : \eta = 0$, only that null

needs to be tested, but as seen below, multiple versions of this bread block will be created when generating a confidence interval.

For the hypothesis test of $H_0 : \eta = 0$, it is sufficient to generate a test statistic

$$\frac{\hat{\eta}}{\sigma(\hat{\eta})} \tag{3.1}$$

by the ratio of $\hat{\eta}$ and its standard error. As in typical OLS settings, it will be a t-statistic, using the degrees of freedom from the first stage model.

The same procedure is used for the intercept, interpretable as the average treatment effect when X is at its mean.

3.2.2 Confidence Intervals

The confidence region will be generated by test inversion, where we iterate over a range of $H_0 : \eta = \eta_0$, and the confidence region is the set of all η_0 such that the hypothesis fails to reject. The test statistic used is

$$t(\hat{\eta}) = \frac{|\hat{\eta} - \eta_0|}{\sigma(\hat{\eta})}, \tag{3.2}$$

rejecting H_0 when

$$t(\hat{\eta}) \geq t^*_{1-\alpha/2}. \tag{3.3}$$

By squaring both sides and re-arranging, we obtain an expression which is quadratic in $\hat{\eta}$.

This gives us two benefits. First, rather than iterating over a range of hypotheses, we can test three arbitrary hypotheses, use them to generate the coefficients of the quadratic curve, and solve the quadratic equation to obtain the bounds of the confidence region.

Secondly, once we obtain the coefficients of the quadratic, we can easily determine

the shape and minimum/maximum of the curve. If the curve is convex, we know the confidence interval must be finite. If the curve is concave and its maximum is positive, the confidence region is disjointly infinite. If the curve is concave and its maximum negative, the confidence region is infinite.[1]

Confidence intervals are obtained by passing a `pbph` object (the result of a call to `pbph`) to `confint`. It extends `confint.lm` in input and output.

In Section 2.7, we recommend only considering the confidence interval if the hypothesis test rejects the null. If the hypothesis test in the `pbph` failed to reject, `confint` returns `(NA, NA)`. This can be overridden by passing `forceDisplayConfInt = TRUE`. If the confidence interval is disjoint, or if `returnShape = TRUE` is passed as an argument, than an additional attribute, `shape`, is returned taking values of either `finite`, `infinite` or `disjoint`.

3.3 Additional Complications

We now show the implementation of two additional complications. First, we will allow the first stage to be a generalized linear model. Secondly, we will allow handling clustered random trials.

3.3.1 PBPH with GLM First Stage

If we do not assume that the error on a response variable is normally distributed, a linear model may not be appropriate. Generalized linear models (GLM), which allow the error to have any exponential family distribution, may be more appropriate. If the user specifies that the first stage model is a GLM, the second stage remains linear. If Y is drawn from a particular distribution, there is no reason to assume that $Y - \hat{Y}_c$ needs to also follow the same distribution. By using a linear second stage, we are

[1] We showed in Section 2.5.4 that a convex curve with positive minimum, corresponding to an empty rejection region, is not obtainable.

examining whether the treatment effect is additive. Additionally, there is a practical consideration. Consider the situation where Y is binary. Despite Y being binary, $\hat{Y}_c$ would almost surely not be, so $Y - \hat{Y}_c$ would no longer even be discrete.

For the full derivation, see Appendix B.1. While the derivation should hold for any GLM with a canonical link, the implementation only allows for a binomial or Poisson first stage model at this time.

The general implementation does not differ; the calculation of the bread includes a term for the estimated variance of each observation, so that the calculation of the bread is

```
crossprod(x, x * vhat)
```

where `vhat` is a vector of estimated variances.

3.3.2 Clustered Standard Errors

Randomization can be performed across clusters instead of individuals. When individual randomizing is infeasible, it may be more useful to randomize by group instead. The target population is divided into mutually exclusive groups. Typically, these groups have some natural definition, such as blocks, schools, cities, etc.[47]

The convenience of clustered random trials is balanced with a loss in precision and power in a cluster experiment compared to simple random assignment with the same number of individuals.[21] Members within clusters are likely to be more homogeneous than those across clusters, introducing an intracluster correlation. Because of this intracluster correlation, the effective sample size of a set of clustered data is diminished, yielding underestimated standard errors.[28]

We do not consider any cluster-level effects, only allowing for the adjustment needed for intracluster correlation.

Sandwich estimators are a common tool to handle the standard errors in clustered

data situations. The calculations of the meat are modified to first sum the estimating functions within each cluster before taking across-cluster variation.[17] To see the full derivation, see Appendix B.2.

We implement this by overloading the `meat` function from the **sandwich** to allow a `cluster` argument. The relevant modifications[2] are

```
psi <- sandwich::estfun(x)
if (!is.null(cluster)) {
  psi <- aggregate(psi, by = list(cluster), FUN = sum)
}
```

Additionally, there is need for a finite sample adjustment of the form

$$\frac{S}{S-1} \cdot \frac{n-1}{n-p}, \tag{3.4}$$

where S is the number of clusters, n is the number of observations and p is the number of parameters in the model.[17] If $S = n$, where each observation is its own cluster, this is equivalent to not using a clustered sampling method. Then (3.4) collapses to $\frac{n}{n-p}$, which is a common degree of freedom adjustment in regression settings [48] and the default in **sandwich** [71].

3.4 Examples

Here we give several examples of implementation of the methodology. Each example uses a data set included in the **pbph** package.

[2] `sandwich` is also modified to pass the `cluster` argument down to `meat`.

3.4.1 PBPH with Linear First Stage

The `eottest` data contains student performance on an exam ("`test`"), the student's class ("`class`", which will be used later in clustering), demographics ("`male`", "`gpa`"), and participation in an after school program ("`afterschool`"). We wish to see whether the after school program is effective, and whether it is more effective on those who are most likely to fail the test in the absence of any intervention.

```
data(eottest)
mod1 <- lm(test ~ gpa + male, data = eottest,
           subset = (afterschool == 0))
```

We fit the first stage model on only the control group, which is defined as a 0 in "`afterschool`". The first stage fits very well, which should always be a goal.

Now, we fit the second stage model using `pbph`.

```
mod2 <- pbph(mod1, treatment = afterschool, data = eottest)
summary(mod2)

##
## Call:
## lm(formula = test - pred ~ treatment + pred, data = newdata,
##     subset = (treatment == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2983 -0.9854 -0.2190  0.9647  3.4119
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
```

```
## treatment   3.3085     0.2592  12.765   <2e-16 ***
## pred       -0.4885     0.1137  -2.265   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 38 degrees of freedom
## Multiple R-squared:   0.91,Adjusted R-squared:  0.9053
## F-statistic: 192.1 on 2 and 38 DF,  p-value: < 2.2e-16
```

The standard error and associated p-value are computed using the PBPH method. We can also obtain a confidence interval,

```
confint(mod2)

##                2.5 %      97.5 %
## treatment  2.7838090  3.83316671
## pred      -0.6528415 -0.08992958

confint(mod2, returnShape = TRUE)

##                2.5 %      97.5 %
## treatment  2.7838090  3.83316671
## pred      -0.6528415 -0.08992958
## attr(,"shape")
## [1] "finite"
```

and optionally return the shape of the confidence interval for reassurance.

3.4.2 PBPH with Logistic Data

Data `salesdata` can be used to test whether a new sales technique is effective in increasing sales. The data contains indicators of successful sales ("`sale`") and whether the new technique was randomly chosen to be used ("`newtechnique`"), and some information about the salesperson ("`experience`" and "`previousales`").

Since the response in binary, the first stage model is a logistic regression model.

```
data(salesdata)
mod1 <- glm(sale ~ experience + previoussales, data = salesdata,
            subset = (newtechnique == 0), family = binomial)
```

Regardless, the second stage is fit the same

```
mod2 <- pbph(mod1, treatment = newtechnique, data = salesdata)
summary(mod2)

##
## Call:
## lm(formula = sale - pred ~ treatment + pred, data = newdata,
##     subset = (treatment == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3167 -0.2551 -0.1908 -0.1570  0.8309
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## treatment -0.20296    0.03878  -5.234 1.66e-07 ***
## pred      -0.82498    0.15642  -2.292   0.0237 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4196 on 117 degrees of freedom
## Multiple R-squared:  0.3477,Adjusted R-squared:  0.3365
## F-statistic: 31.18 on 2 and 117 DF,  p-value: 1.402e-11
```

The conclusion is that the new sales technique lowers the odds of a resultant sale, but that effect is strongest (most negative) on those most likely to have made the sale using the old technique. In other words, the new technique may assist poor performing or newer salespeople, but those with a proven track record are unlikely to be assisted.

3.4.3 Clustered Data

We return to the student test data. The data can be thought of a clustered random trial, where classrooms were assigned to the after school program instead of individual students. Very little modification is needed to enable this.

```
mod1 <- lm(test ~ gpa + male, data = eottest,
           subset = (afterschool == 0))
mod2 <- pbph(mod1, treatment = afterschool, data = eottest,
             cluster = class)
summary(mod2)

##
## Call:
## lm(formula = test - pred ~ treatment + pred, data = newdata,
##     subset = (treatment == 1))
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2983 -0.9854 -0.2190  0.9647  3.4119
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## treatment   3.3085     0.1325  24.979  < 2e-16 ***
## pred       -0.4885     0.0777  -4.446  7.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 38 degrees of freedom
## Multiple R-squared:   0.91,Adjusted R-squared:  0.9053
## F-statistic: 192.1 on 2 and 38 DF,  p-value: < 2.2e-16
```

```
confint(mod2)
```

```
##                2.5 %     97.5 %
## treatment  3.0403521  3.5766236
## pred      -0.6302728 -0.3086221
```

Note that the estimate of the interaction coefficient does not change, but we properly compute the standard error.

3.5 Simulations

We include some simulation results to show the validity of these extensions.

3.5.1 Logistic first stage

To examine the generalized linear model first stage extension, we focus on a common variation, that of a logistic first stage.

3.5.1.1 Data Generation

The covariates X are generated randomly from $N(0,1)$, $X \in \mathbb{R}_{n\times q}$. We use $n = 100$ and $n = 1{,}000$ for smaller and larger sample situations. We use $q = 7$ and $q = 17$ for $n = 100$ and $n = 1{,}000$ respectively, due to the rule of thumb we developed in Section 2.6.2.2. q describes merely the dimensions of the generated X, and it can (and often will be the case) that the response Y is generated by a data generating matrix of dimension $n \times p$, which is a subset of X, such that $p < q$. This distinction is why we use q to represent the dimension of X and p to represent the dimension of the data generating matrix.

β_c in the first stage model are drawn from $N(0,1)$, with some $q-p$ of the β_c forced to 0 to add some noise.

The choice of second stage model parameters τ and η require a bit more finesse than in the linear first stage case. In those cases, the only restriction on these parameters in a simulation was that $\eta \in (-1,2)$, those values being chosen to restrict attention to models where the relationship between X and Y_c and between X and Y_t are similar.

However, in the cases where the response in the not normally distributed and the first stage is a generalized linear model, there are additional restrictions upon η and τ. Unlike the linear-linear restrictions (where an estimate of η outside of $(-1,2)$ might indicate that there are additional model complexities that the method does not address), these restrictions are purely mathematical.

Consider the case where Y is binary and the first stage model is logistic. In this setting, $\hat{Y}_c = \text{logit}^{-1}(X\hat{\beta}_c)$ and the left hand side of the second stage model, $Y_t - \hat{Y}_c$,

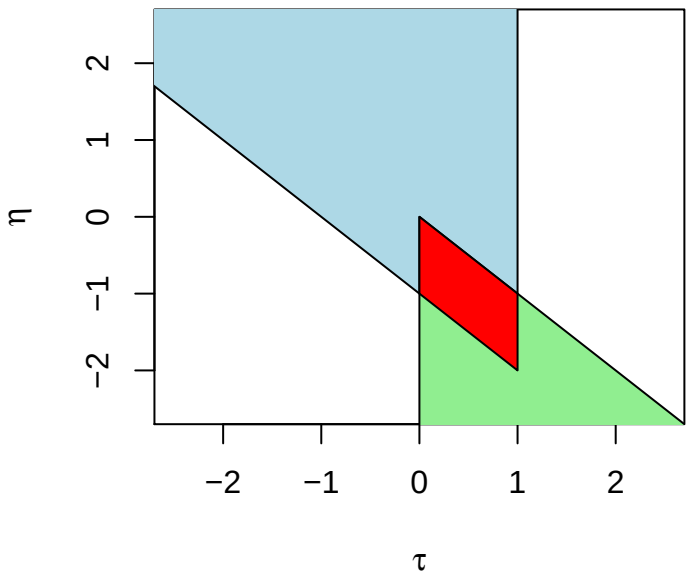

Figure 3.1: Possible values of η and τ. The light green and light blue represent the two regions described in the text whose union covers mapping of some values of $\text{logit}^{-1}(X\hat{\beta}_c)$ to $[-1, 1]$, and the red represents their intersection which maps all values of $\text{logit}^{-1}(X\hat{\beta}_c)$ to $[-1, 1]$.

is restricted to $[-1, 1]$. For example, we cannot have that $\tau = 1.5$ and $\eta > 0$, as then for all values of $\hat{Y}_c \in [0, 1]$, the right hand side never maps to $[-1, 1]$. If $\tau = 0.5$ and $\eta = -0.4$, then all values of $\hat{Y}_c \in [0, 1]$ maps the right hand side into $[-1, 1]$. There are some cases with partial successful matching, for example, if $\tau = -0.25$ and $\eta = 0.5$, then $\hat{Y}_c \in [0, 0.5)$ does not map into $[-1, 1]$ while $\hat{Y}_c \in [0.5, 1]$ does.

There are two regions of interest in defining, in the logistic case, feasible values of the parameters, $\{\tau \geq 0, \eta \leq -\tau\}$ and $\{\tau \leq 1, \eta \geq -1 - \tau\}$. The union of those two regions corresponds to values of η and τ where some values of $\hat{Y}_c \in [0, 1]$ map to $[-1, 1]$. The intersection of those regions corresponds to values of η and τ where all values of $\hat{Y}_c \in [0, 1]$ map to $[-1, 1]$. This is visually represented in Figure 3.1.

Finally, we generate success probabilities for each individual using

$$\rho_i = \text{logit}^{-1}(-X_i\beta_c) + \tau \times Z + \eta \times Z \times \text{logit}^{-1}(-X_i\beta_c). \tag{3.5}$$

In the control group, $Z = 0$ and $\rho_i = \text{logit}^{-1}(X_i\beta_c)$. When $Z = 1$ in the treatment group, we have the additional additive effects. We truncate values of ρ_i above 1 or below 0. From this, we generate Y_i from Bern(ρ_i).

3.5.1.2 Simulation Results

We generate values of η and τ which fall within the regions described above, and run 1,000 repetitions each, generating a coverage percentage. These coverage percentages are plotted in Figure 3.2 for $n = 100$ and Figure 3.3 for $n = 1{,}000$.

The red area, where values of τ and η map all values of $\text{logit}^{-1}(X\beta_c)$ into $[-1, 1]$, shows proper coverage which is slightly conservative. However, the blue and green areas, where values of the parameters map some values of $\text{logit}^{-1}(X\beta_c)$ into $[-1, 1]$, shows poor coverage, a problem which is exacerbated with the larger n. This suggests the need to be very careful if $\hat{\eta}$ and $\hat{\tau}$ fall outside of the red area, as the type I error will be large.

3.5.2 Clusters

Data generation follows Section 3.5.1.1 with a few modifications. Following the notation of Section B.2, let each of n observations belong to exactly one of S clusters and n_s observations belong to cluster s. Due to the intracluster correlation discussed in Section B.2, we require a larger sample size to obtain similar power to the non-cluster version.[31] In practice, we will use $S = 10$ and 100 and $n = 1{,}000$ and 4,000, examining all four pairings, to see the difference in effect of the size of S versus the effect of the size of n. We randomly assign observations to a cluster with equal

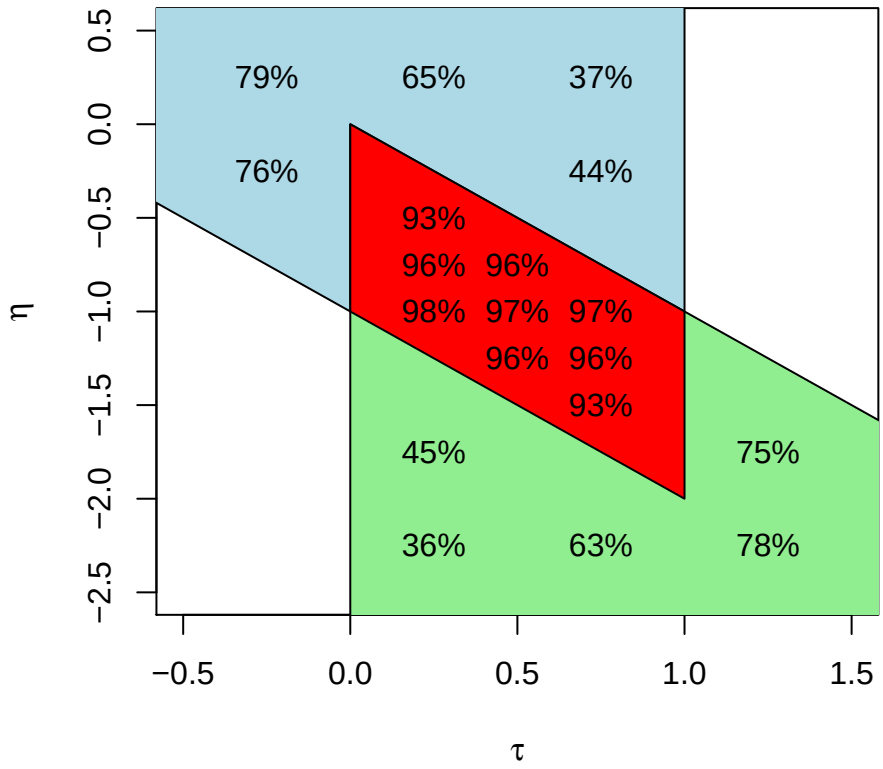

Figure 3.2: Coverage over different values of η and τ for $n = 100$. The light green and light blue represent the two regions described in the text whose union covers mapping of some values of $\text{logit}^{-1}(X\hat{\beta}_c)$ to $[-1, 1]$, and the red represents their intersection which maps all values of $\text{logit}^{-1}(X\hat{\beta}_c)$ to $[-1, 1]$.

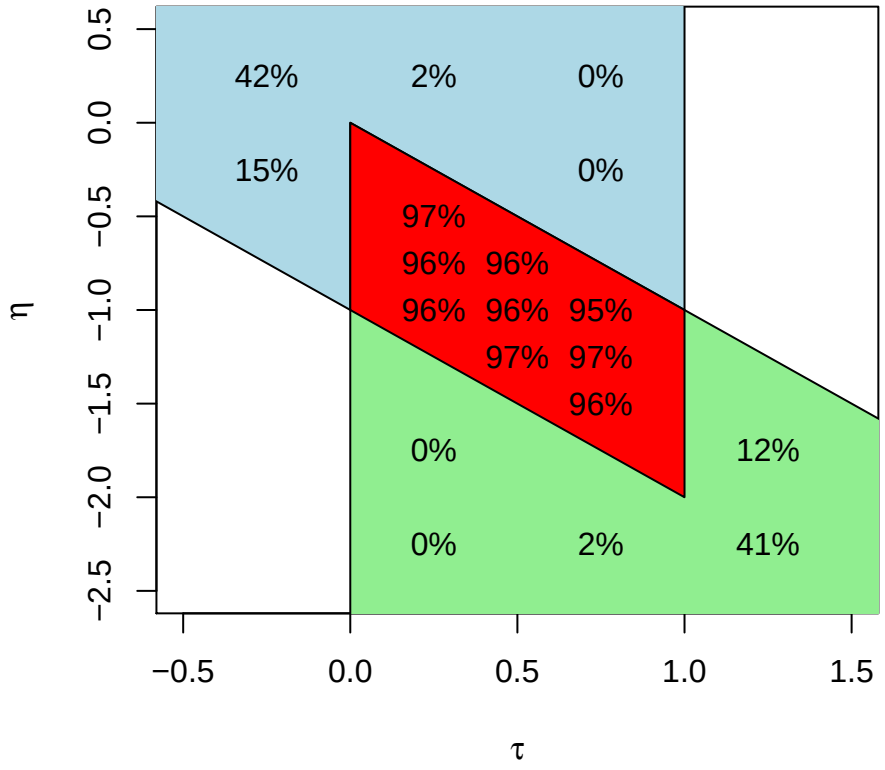

Figure 3.3: Coverage over different values of η and τ for n =1,000. Coverage is less conservative in the red area, and performs much worse in the outlying regions.

probability, so that $\mathbb{E}(n_s) = n/S$. With equal probability of assigning each cluster to treatment or control, we have $\mathbb{E}\left(\sum Z_i\right) = n/2$, so we use $q = 13$ and 22 for $n = 1{,}000$ and 4,000 respectively, again following the rule of thumb from Section 2.6.2.2. We do not include any cluster-level effects.

For choices of η and τ, the addition of clusters does not affect results from the most basic case, so we merely limit η to $(-1, 2)$.

3.5.2.1 Cluster Simulation Results

For each combination of $(S, n) \in \{(10, 100) \times (1000, 4000)\}$, we run 1,000 replications as described above and examine coverage. The results are plotted in Figure 3.4. With the largest configuration, we see proper 95% coverage. As S or n decrease, the coverage drops. The effect of S decreasing is much more substantial. Since shrinking n has a very minor impact, we can interpolate implying that the size of $\mathbb{E}(n_s)$ also plays a small role.

An additional note is that because n is so large in these simulations, the bias we observed in the linear variation of the corrected PBPH vanished.

3.5.3 Revisiting Giné et al. [29] with Clusters

We revisit Giné et al. [29], the paper which motivated this work (see Section 2.3). The authors performed an uncorrected PBPH method to determine whether fingerprinting farmers applying for loans in rural Malawi improved repayment rates, and whether the improved repayment was greatest for those most likely to default. The paper found affirmative answers to both questions.

In Section 2.6.5, we re-analyzed their results using the corrected PBPH and confirmed their results, with the caveat that the confidence interval we generated almost covered -1. An interaction coefficient of -1 would indicate that there is no relationship between the potential responses.

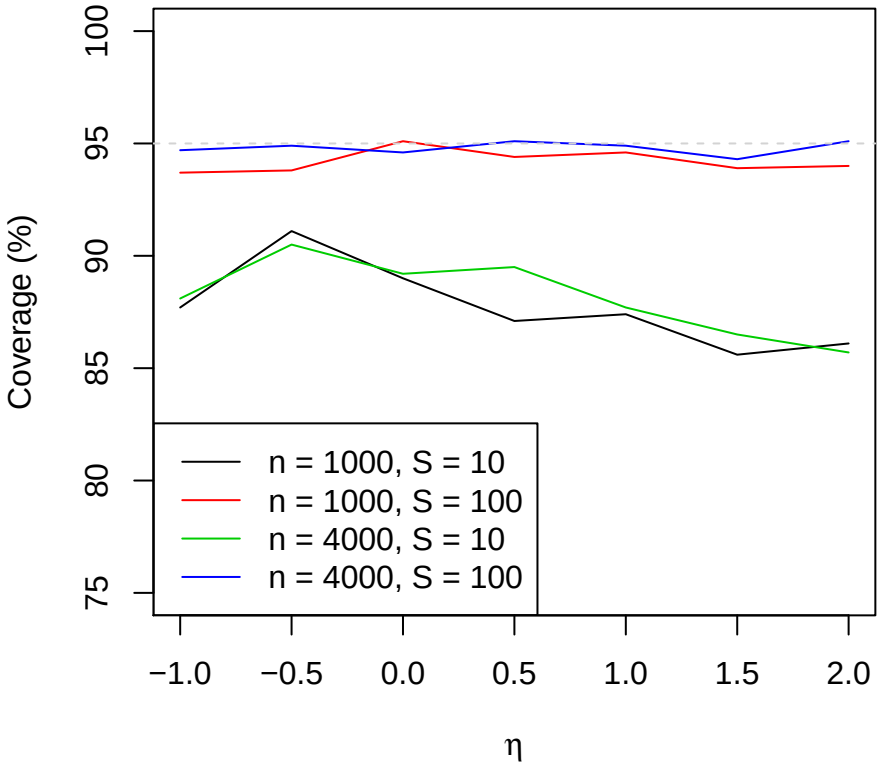

Figure 3.4: Coverage over η with S clusters and sample size n. As either n or S increases, coverage increases, though the effect of S is more extreme.

In Giné et al. [29], the unit of randomization was not farmer but "club", a collection of farmers who share risk. We ignored this complication in Section 2.6.5, but with the addition of allowing clustered randomized trials, we can include the clubs. The results are shown in Table 3.1.

	Estimate	S.E.	Confidence Interval
Uncorrected	-0.896	0.043	(-0.980, -0.812)
Corrected	-0.896	0.054	(-0.998, -0.781)
Corrected w/ Clusters	-0.896	0.109	(-1.110, -0.635)

Table 3.1: Comparison of uncorrected and corrected confidence intervals, adjusting for clustered randomized trials.

With proper handling of the clubs, we now see a confidence interval that does cover -1. This suggests the need to revisit results, as no relationship between potential responses is a negative result.

3.6 Conclusion

We demonstrate the implementation of the PBPH methodology by introducing the `pbph` package. The package enables users to easily fit the second stage model which will correct the standard error to account for the two-stage modeling setting.

Following this, we demonstrated two embellishments on the methodology which the working statistician may encounter. First, we generalize the method to account for non-normal residuals by allowing the first stage model to be a generalized linear model, for example logistic regression for binary data. Additionally, we allow the analysis of data generated via clustered random trials, by correctly computing the standard errors accounting for the clusters.

Simulations allowed the exploration of nuances of these methods. We showed the restrictions on the coefficients that exist with a logistic first stage, as well as discussed some heuristics for the sample sizes needed to obtain adequate coverage with clustered

data. Finally, we showed the implementation of both these embellishments on Giné et al. [29].

Together, these enhancements to the PBPH methodology offer a nice set of flexibility to the working statistician which shows the strength of our methodology, though of course, further enhancements are possible.

CHAPTER IV

Enabling Linear Treatment Effects with a Binary Response

4.1 Introduction

When examining the treatment effect with a binary response, typically used models are general linear models with a logistic link function or conditional logistic regression. While there are several reasons to prefer these models over linear regression models [22], one side effect is that the treatment effect is forced to be linear on the logit scale. In other words, the treatment effect is multiplicative on the probability scale - the scale which usually has the easiest interpretation.

We examine methods to test whether a treatment effect on a binary response is linear on the logit scale or linear on the probability scale. Rather than fitting a single linear model, which would additionally force all predictors to be linear on the probability scale, we use a two stage least squares procedure, where the first stage is logistic and the second is linear. The second stage contains only the effect of treatment, while the first contains all other predictors.

We further show that by comparing the linear second stage vs a logistic second stage, choosing the model which minimize the expected risk function based upon logistic loss can help determine which linearity scale more closely captures the treatment

effect.

Another common feature of treatment effect analysis is stratification of the observations, for example through matching. In these settings, conditional logistic regression is a common approach (given the inconsistency of ordinary logistic regression with stratum fixed effects [2]). We first show how by fitting two slightly different two-stage logistic regression models, we can examine whether there is evidence that the treatment effect may be linear on the probability scale. If that evidence exists – and in some cases even if it does not – then a similar two-stage approach can be taken with a logistic first stage and a linear second stage. We ultimately recommend using weighting in the second stage to account for the strata as opposed to conditional logistic regression or fixed effects.

Section 4.2 will examine the setting with no strata, examining the difference between the scales in detail in Section 4.2.3 and choosing the model in Section 4.2.4. Section 4.3 shows simulation results.

In Section 4.4 we turn to the setting with strata, showing both using conditional logistic regression to gain evidence towards what scale the treatment effect is linear on in Section 4.4.2, and the modifications to the two-stage procedure to account for the strata in Section 4.4.3.

Finally, we apply our results to Gurm, Hosman, Share, Moscucci, and Hansen [32] in Section 4.5.

4.2 Linear vs Logistic

4.2.1 Logistic Regression

In general, there are numerous valid reasons to prefer a logistic regression model over a linear model when Y is binary. For example, in linear regression, the restriction that $\hat{Y} \in [0, 1]$ is not enforced, extrapolation is more hazardous than usual, and we

know the response distribution is non-normal so the residuals will be incorrectly modeled. These restrictions are discussed in length in numerous sources, for example Agresti [2] or Cox and Snell [22].

Logistic regression models address these concerns and enable a more robust analysis of the data. It is important to note that the methods we are proposing are not a framework for considering a logistic vs linear model in a general setting; rather we are restricting ourselves to the setting where the predictor of interest is Z, the treatment indicator. There can be other predictors X, but they must be modeled in a first stage to that the relationship between Y and X remains firmly in the logistic framework.

A second benefit of the two-stage approach is separating the tasks of modeling the relationship between the response and its predictors from the task of modeling the treatment effect. A one-stage model which includes both X and Z must address both issues simultaneously.

4.2.2 Loss and Risk Functions

Regression can be thought of as a process to find a function $f(X)$ which minimizes some risk function for the prediction error from predicting Y with $\hat{Y} = f(X)$. For example, in linear regression, $f(X) = X\beta$. The risk function is the expected value of a loss function, which is any function $L(Y, f(X))$ which has properties

$$\begin{aligned} L(Y, Y) &= 0, \\ L(Y, f(X)) &\geq 0. \end{aligned} \tag{4.1}$$

Our treatment of loss and risk is somewhat informal; more formal discussion appear in literature such as statistical decision theory (e.g., Keener [43, Ch. 11], Hastie, Tibshirani, and Friedman [34, Ch. 2, 7]) and classification problems (e.g., Bartlett, Jordan, and McAuliffe [8], Freund, Schapire, Singer, and Warmuth [27]).

There are different choices for the loss function over which to optimize the choices

of $f(x)$. Linear regression is usually solved with the least squares method which uses the quadratic loss function,

$$(Y_i - f(X_i))^2, \tag{4.2}$$

whose risk (expected loss) can be estimated by

$$\frac{1}{n}\sum_i (Y_i - f(X_i))^2. \tag{4.3}$$

It is possible to fit linear regression with other loss functions; another common example is least absolute difference, which can be more robust than least squares, but admits multiple solutions.[43, 44]

For binary outcomes, there are many choices of loss functions motivated by classification problems such as 0/1 loss, hinge loss, or boosting loss.[15] Logistic regression performs by minimizing the logistic loss function,

$$-Y_i \log(f(X_i)) - (1 - Y_i)\log(1 - f(X_i)), \tag{4.4}$$

with a similarly defined estimated risk.

In addition to minimizing these loss functions to fit the regression models, the loss functions can be used for model selection. Consider two competing regression model, the first with predictors $X^{(1)}$ and the second with predictors $X^{(2)}$. Then we choose the first model only if

$$\sum_i (Y_i - X_i^{(1)}\hat{\beta}^{(1)})^2 < \sum_i (Y_i - X_i^{(2)}\hat{\beta}^{(2)})^2. \tag{4.5}$$

However, if Y is binary and we are comparing a linear and logistic model, the decision criterion is not as clear, as the loss function fitting each model is different. As we stated earlier, we are not offering a general solution. However, in the limited setting where our goal is to determine whether a treatment effect is linear on the logit

scale or linear on the probability scale, we will present evidence from simulations that using the logistic loss function, (4.4), is superior to the quadratic loss function in the sense that it more commonly chooses the model which is based upon the data-generating model.

4.2.3 Treatment on Probability or Logit Scale

To see the difference of a treatment effect on the two scales, lets take a simple example. This toy example will be represented in a one-stage model for ease of understanding, while our method relies on the two-stage variation.

Let there be binary response Y, treatment indicator Z and some grouping variable G with two categories. Say the true conditional probabilities are

$$P(Y = 1|Z = 0, G = 1) = .05, \tag{4.6}$$

$$P(Y = 1|Z = 0, G = 2) = .50, \tag{4.7}$$

$$P(Y = 1|Z = 1, G = 1) = .15. \tag{4.8}$$

The remaining true conditional probability, $P(Y = 1|Z = 1, G = 2)$, will obviously have different values depending on the true model. If the true model is linear,

$$P(Y|Z, G) = \alpha_1 1_{G=1} + \alpha_2 1_{G=2} + Z\tau, \tag{4.9}$$

then we have that

$$\begin{aligned} \tau = P(Y = 1|Z = 1, G = 1) - P(Y = 1|Z = 0, G = 1) = \\ P(Y = 1|Z = 1, G = 2) - P(Y = 1|Z = 0, G = 2). \end{aligned} \tag{4.10}$$

In other words, the effect of the treatment on the probability scale is constant

across the groups of G. Therefore,

$$P(Y = 1|Z = 1, G = 2) = .60. \tag{4.11}$$

On the other hand, if the true model is linear on the logit scale (i.e. a logistic regression model),

$$\text{logit}\left(P(Y|Z, G)\right) = \alpha_1 1_{G=1} + \alpha_2 1_{G=2} + Z\tau, \tag{4.12}$$

then (4.10) no longer holds as linearity of the treatment effect exists only on the logit scale. In this setting, the remaining conditional probability would be

$$P(Y = 1|Z = 1, G = 2) \approx .77. \tag{4.13}$$

A visual representation of this is included in Figure 4.1.

4.2.4 Model comparison

In the simple set-up discussed in Section 4.2.3 where the only additional predictor is binary, we can choose between (4.9) and (4.12) simply by comparing the slope defined by the observed responses to treatment versus the observed responses to controls. If there is a significant difference between them, that can be considered evidence that linear model on the probability scale, (4.9), is unlikely. However, if the predictors are of higher dimension, the analysis gets much more complex. We will address this comparison by minimizing an expected risk function.

The use of risk functions in model selection is not novel, for example using the mean-squared error in cross-validation.[56] Let Y be the observed response that we are attempting to predict, and let $\hat{Y} = f(X, Z)$ be some prediction obtained by a regression model. In our framework, this model can be logistic or linear; if it is linear,

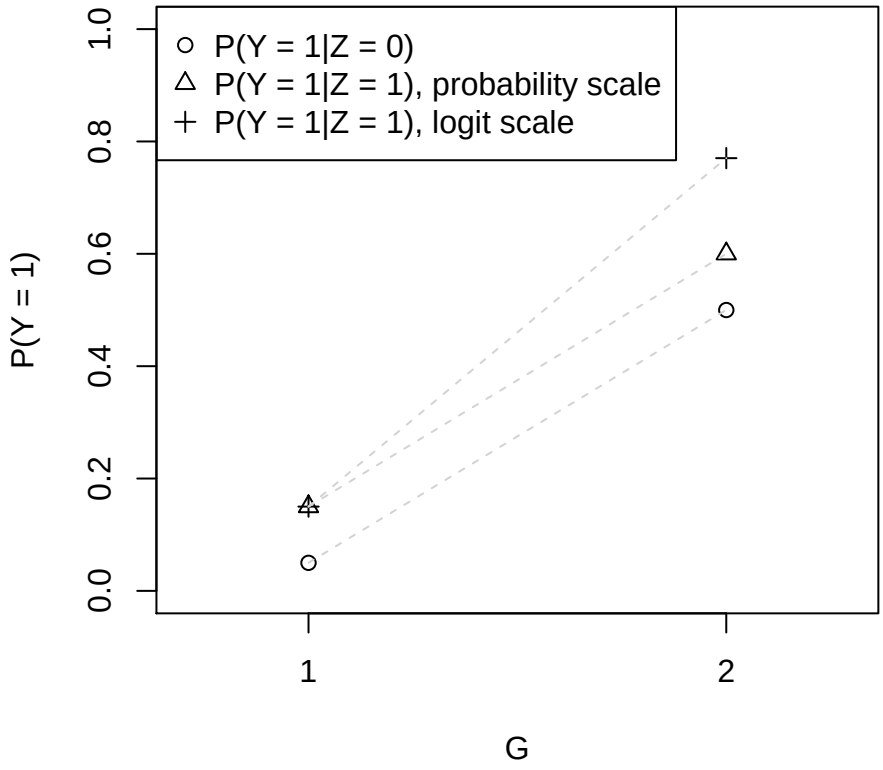

Figure 4.1: A demonstration of a linear treatment effect on the probability and logit scales.

replace values of $\hat{Y}$ outside of $[0,1]$ with the closer of $\{0,1\}$, to mimic the general understanding of out-of-range predictions. We can estimate the overall risk by the average risk in the sample, for example using the quadratic loss function (4.2),

$$\hat{R}_{\text{quad}}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y-\hat{Y})^2. \tag{4.14}$$

This is known as the predictive risk.[25] Because we are restricting the response $Y \in \{0,1\}$, the quadratic loss simplifies to

$$Y(1-\hat{Y})^2 + (1-Y)\hat{Y}^2. \tag{4.15}$$

The risk using the logistic loss function, (4.4), can be similarly defined as

$$\hat{R}_{\log}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\Big(-Y\log(\hat{Y}) - (1-Y)\log(1-\hat{Y})\Big). \tag{4.16}$$

As we show below in Section 4.3.2, if we choose between a linear or logistic second stage model which minimizes $\hat{R}_{\log}(Y,\hat{Y})$, we can gain evidence as to on which scale the linearity of the treatment effect is more closely aligned.

4.3 Simulations

4.3.1 Data Generation

Let there be some predictor X of response $Y \in \{0,1\}$. We have $Z \in \{0,1\}$ representing membership in a control and treatment group respectively. The goal is to fit a two step model, where the first step is a logistic model fit only on the control group,

$$\text{logit}\left(\mathbb{E}(Y_c|X)\right) = \alpha_0 + \beta_0 X. \tag{4.17}$$

Then $\hat{Y}_c$ is the predicted response to control. In the second stage, we wish to

determine whether the effect of treatment is additive (on the probability scale) or multiplicative (additive on the logit scale). The two comparison models are

$$\mathbb{E}(Y|Z, \hat{Y}_c) = \beta_1 Z + \hat{Y}_c, \tag{4.18}$$

for additive in the probability scale, and

$$\text{logit}\left(\mathbb{E}(Y|Z, \hat{Y}_c)\right) = \beta_2 Z + \text{logit}\left(\hat{Y}_c\right), \tag{4.19}$$

for additive on the logit scale.

We will draw $X \sim N(0, 1)$. If β_0 is close to 0, then $\hat{Y}_c$ will have little variation, and differentiating between (4.18) and (4.19) is difficult. Additionally, differentiating between models will be difficult if the treatment effect (β_1 or β_2) is small. To visualize this, see Figure 4.2. As β_2 decreases, the logistic model fit becomes closer to linear, and differentiating the two models is difficult. However, as β_2 increases, the difference between the models is easier to detect.

4.3.2 Results

We compare the two risk functions, (4.14) using the quadratic loss function and (4.16) using the logistic loss. We then define the decision criterion to choose (4.18) if the risk associated with (4.18) is smaller than (4.19).

The results across varying values of β_1 and β_2 are in Figure 4.3.

As we can see from the results, the logistic risk function outperforms the quadratic risk function in choosing the correct scale for the treatment effect. Therefore we recommend using a risk function with logistic loss.

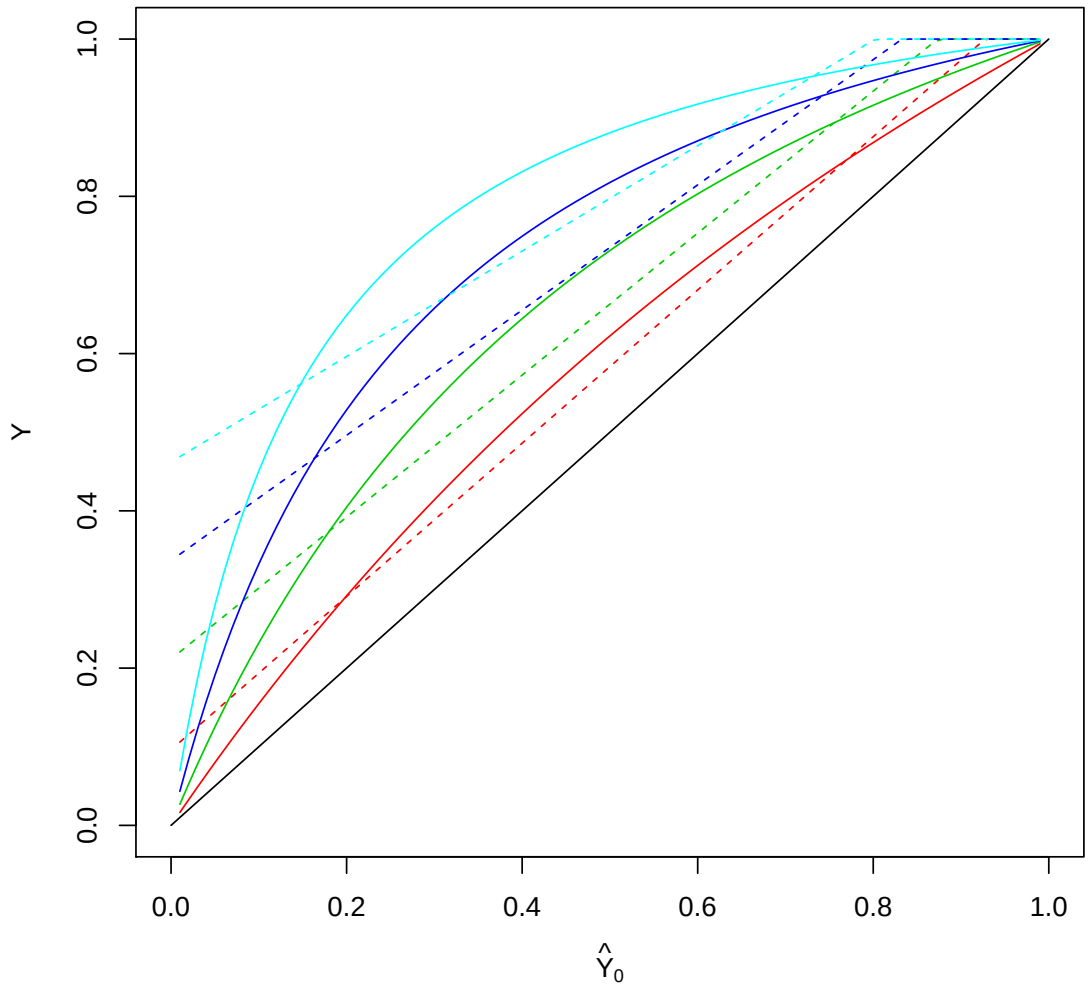

Figure 4.2: Comparison of second stage linear vs logistic model fits. The solid black line is $Z = 0$. Solid lines are from (4.19) with $\beta_2 = .5, 1, 1.5$ and 2 as the lines get further from the black line. The dashed lines are from (4.18) fit upon the logistic fitted values. As the β increases, the lines become easier to distinguish.

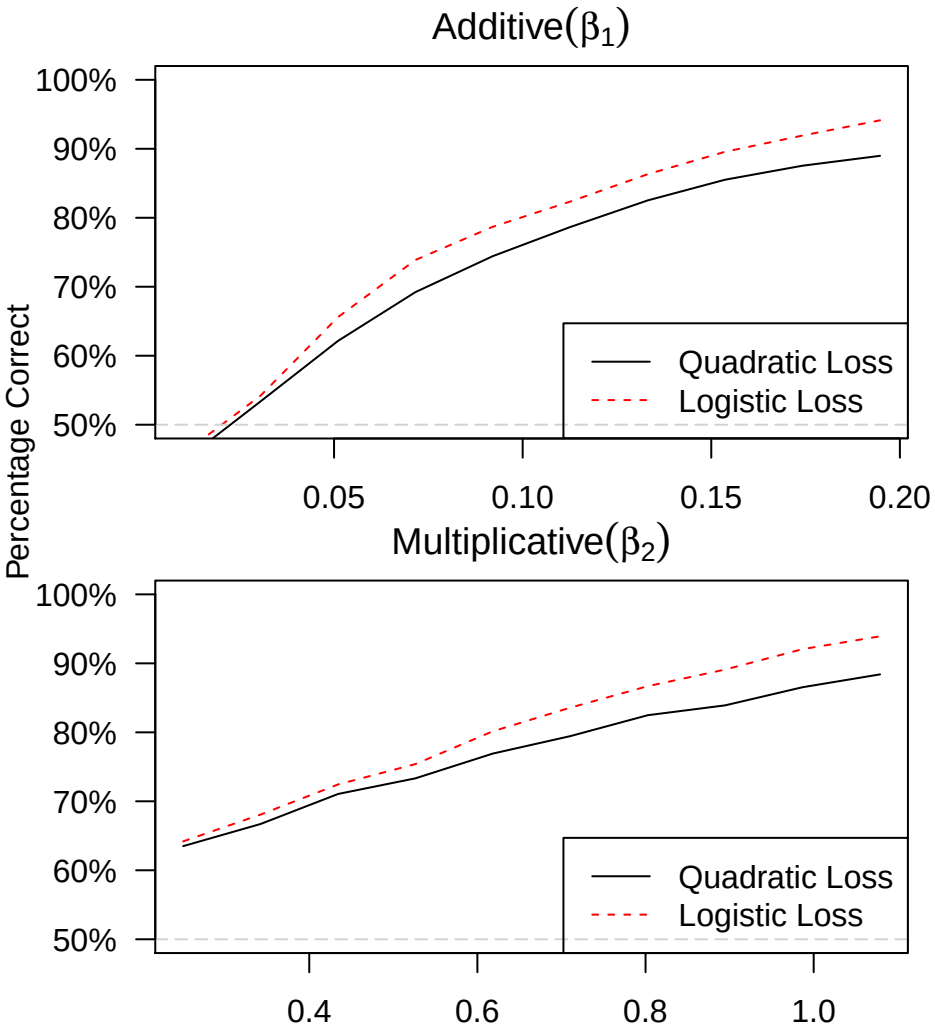

Figure 4.3: Performance of risk functions using quadratic loss and logistic loss. As the effect size increases, both risk functions perform better at choosing the correct model, though the logistic risk function always outperforms the quadratic risk function.

4.4 Linear vs Logistic, with Stratification

4.4.1 Conditional Logistic Regression

Conditional logistic regression is a modification to logistic regression which enables controlling for parameters in the model without needing to estimate coefficients for them. The likelihood is evaluated conditional on realized values of sufficient statistics corresponding to these unwanted parameters, and the resulting quasi-likelihood function is then maximized to find estimates for the remaining parameters.

Let our response be $Y \in \{0,1\}_n$. Let $X \in \mathbb{R}_{n\times p}$ be a set of predictors including a constant column for the intercept, and let $U \in \mathbb{R}_{n\times q}$ be a set of unwanted predictors that need to still be controlled for. These U can generally be any set of predictors, but in this context we will consider them to be fixed effects for strata or matched sets. The logistic model would be

$$\text{logit}\left(\mathbb{E}(Y|X,U)\right) = X\beta + U\gamma, \tag{4.20}$$

with $\beta \in R_p$ and $\gamma \in R_q$.

If $n >> p+q$, this model is sufficient; and we can maximize the likelihood of (β,γ) to obtain $(\hat{\beta},\hat{\gamma})$.

However, if q is large, we run into issues. If $n \leq p+q$, the model is under-specified. In general, as $p+q$ increases relative to n, the performance of the maximum likelihood solution is poor.[2, Ch. 6]

If γ is not of interest, we can condition the likelihood of (β,γ) on the sufficient statistics for U. The conditional likelihood lacks dependence on γ, but can otherwise be maximized in the same fashion to obtain $\hat{\beta}'$. See Agresti [2] or Hosmer and Lemeshow [38], amongst others, for a fuller discussion of conditional logistic models.

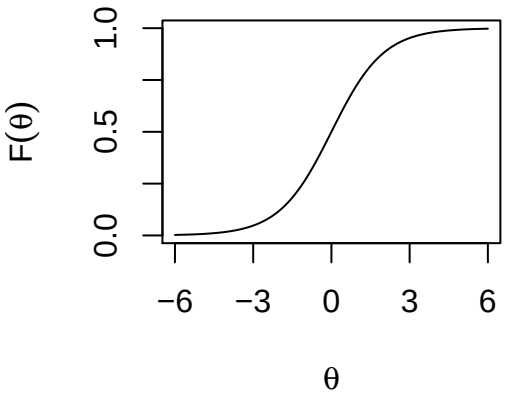

Figure 4.4: F is the inverse logit function.

4.4.2 Evidence for Linear in Probability

The model for a conditional logistic regression approach where matched sets provide the strata can be expressed by

$$\text{logit}\left(\mathbb{E}(Y|X,Z,C)\right) = X\beta + Z\tau + C\gamma, \tag{4.21}$$

where Y is a binary response, X is some design matrix of dimension $n \times p$ including a column of 1's for the intercept, C is the set of indicators for strata membership and Z is treatment status. As described in the previous section, by conditioning on the sufficient statistics for the strata membership, we obtain estimates for (β, τ) without estimating γ.

In this setting, the effect of Z is linear only in the logit scale, and multiplicative in the probability scale. To visualize this, consider Figure 4.4, which plots the relationship between θ and its inverse logit, $F(\theta)$. If θ is near 0, an increase of 1 on the logit scale is equivalent to a shift about 0.23. If θ is near 3, an increase of 1 on the logit scale is equivalent to only a shift of about 0.03.

If the treatment effect is multiplicative in the probability scale (additive on the logit scale), then this is acceptable. However, we may ask whether the treatment effect may be additive on the probability scale. To attempt to gain some evidence towards this end, we begin with conditional logistic model, under the assumption that researchers faced with stratified data and binary response will be very likely to use it for their first analysis.

Let $\bar{Y}_s$ be the average response (or proportion of 1 responses) in strata s. $\bar{Y}_s$ is a natural estimate of $\mathbb{P}(Y = 1|S = s)$. We have that $F'(\theta) = F(\theta)\,(1 - F(\theta))$. For observation i in strata s, let

$$\lambda_{is} = 1/\left(\bar{Y}_s(1 - \bar{Y}_s)\right), \tag{4.22}$$

so that λ_{is} decreases as $\bar{Y}_s$ moves towards .5. Multiplying Z with λ up-weights the treatment effect in strata with $\bar{Y}_s$ closer to 0 or 1, roughly rendering the effect from linear on the logit scale to linear on the probability scale. (Note that $\lambda_{is} = \lambda_{jt}$ if $s = t$, e.g. that i and j belong to the same strata.) Let $\lambda_{is} = 0$ if $\bar{Y}_s \in \{0, 1\}$ as these strata offer no within strata information. We define a new conditional logistic model,

$$\text{logit}\left(\mathbb{E}(Y|X, Z\lambda, C)\right) = X\beta + (Z\lambda)\tau' + C\gamma. \tag{4.23}$$

However, we have entered circular logic, as $\bar{Y}_s$ is what we are trying to avoid estimating by using conditional logistic regression. Because we are assuming the strata are created via matched samples (as opposed to another common use of strata, with a stratifying variable such as gender), the number of strata increases with the sample size, so we gain no further information about any individual $\bar{Y}_s$ as the sample size increases.

If instead of conditional logistic regression, we used two logistic stages, we could introduce the $(Z\lambda)$ term and compare. Assume for the moment that no adjustment

for the strata were required (we will relax this assumption in the following section).

We fit a first stage on only the control group members where $Z_i = 0$ of

$$\text{logit}\left(\mathbb{E}(Y_c|X)\right) = X\beta, \tag{4.24}$$

and then two different second stage models on the entire data set, after getting the predicted response in the absence of treatment on the logit scale, $\text{logit}(\hat{Y}_c)$,

$$\text{logit}\left(\mathbb{E}(Y|Z, \hat{Y}_c)\right) = Z\tau + \text{logit}(\hat{Y}_c), \tag{4.25}$$

$$\text{logit}\left(\mathbb{E}(Y|Z\lambda, \hat{Y}_c)\right) = (Z\lambda)\tau + \text{logit}(\hat{Y}_c). \tag{4.26}$$

Note that there is no coefficient on the predicted response in the absence of treatment, which we will refer to as an offset.

The treatment effect is up-weighted as $\bar{Y}_s$ moves away from .5. Therefore, equal increases on the probability scale will be more closely equivalent on the logit scale as well. If (4.25) outperform (4.26) in some sense (for example, minimizing the risk function associated with logistic loss as discussed in Section 4.3.2), then this is evidence that the treatment effect is linear on the logit scale, and the conditional logistic model is sufficient.

On the other hand, if (4.26) outperforms (4.25), then the treatment effect may be linear on the probability scale, and our two-stage approach can address that.

4.4.3 Modeling Linear Treatment Effect with Stratification

If, after following the recommendation in the previous section, there is evidence that the treatment effect is linear on the probability scale, we can examine the treatment effect more precisely using a linear second stage model. The first stage remains logistic, as it would be preferable to model the relationship between the binary response and the predictors X on the logit scale, regardless of the effect of treatment.

In the first model, as we have been doing, we fit

$$\text{logit}\left(\mathbb{E}(Y_c|X)\right) = X\beta \tag{4.27}$$

amongst the control group only where $Z_i = 0$, to predict response in the absence of treatment. The second stage model is now linear, using the predicted values $\hat{Y}_c = \text{logit}^{-1}\left(\hat{\beta}X\right)$ as an offset.

$$\mathbb{E}\left(Y\middle|Z, \hat{Y}_c\right) = Z\tau + \hat{Y}_c. \tag{4.28}$$

τ is the estimated effect of treatment on the probability scale.

We now consider adjustments to the model to account for the stratification. To start, we can include fixed strata effects,

$$\mathbb{E}\left(Y\middle|Z, S, \hat{Y}_c\right) = Z\tau_f + S\kappa_f + \hat{Y}_c, \tag{4.29}$$

where S is a matrix of indicators of strata membership. This is a very straightforward model to fit, and enables discussion of strata level effects.

As an alternative, consider weighting. Let S_i be the strata membership of observation i. Define δ,

$$\delta_i = \begin{cases} \frac{\sum_{j:S_j=S_i} Z_j}{\sum_{j:S_j=S_i} 1}, & Z_i = 1, \\ \frac{\sum_{j:S_j=S_i} (1-Z_j)}{\sum_{j:S_j=S_i} 1}, & Z_i = 0. \end{cases} \tag{4.30}$$

That is, δ_i is the proportion of observations in the strata which observation i belongs to which have the same treatment status. Note that if observations i and j have $S_i = S_j$ and $Z_i = Z_j$, then $\delta_i = \delta_j$. When strata are sets created via matching, Rosenbaum [57] argues that the unconditional probability of $Z_i = 1$ within strata is constant, and extends this to the probability conditional on the size and structure of the matches, a claim which requires strong ignorability.[57]

In other words, δ_i is an estimate for the probability that an observation was assigned to treatment status Z_i in strata S_i.

Now, let $w_i = \delta_i^{-1} / \sum_j \delta_j^{-1}$ be the normalized inverse and treated as weights, then (4.28) (adding a subscript of w to τ to distinguish) becomes a weighted least square model. The estimate for τ_w is a Hajek-style estimate of the treatment effect (see Appendix C.1 for derivation),

$$\hat{\tau}_w = \frac{\sum_i w_i Z_i (Y_i - Y_{ci})}{\sum_i w_i Z_i}. \tag{4.31}$$

In a true Hajek-style estimate, w_i would represent estimated probability of inclusion in the sample.[9]

If the effect of treatment is constant across strata, then $\hat{\tau}_f$ and $\hat{\tau}_w$ are both estimates of that constant treatment effect. However, if the treatment effect is not constant across strata, then $\hat{\tau}_f$ from the fixed effects model will instead of estimating some weighted average of the strata-specific treatment effects. A benefit of the weighted approach is that $\hat{\tau}_w$ remains a consistent estimate of an average treatment effect regardless of whether the treatment effect is constant.

As is common in two-stage least squares procedures, special consideration must be given to standard error attached to $\hat{\tau}_w$. If the two stage least squares is done manually, the standard error associated with the second stage which utilizes traditional one-stage fitting procedures with $\hat{Y}_c$ in place of Y_c will be negatively biased, as they do not consider the measurement error on $\hat{Y}_c$ introduced by the first stage.[68] This can be addressed with sandwich estimators.

We can reframe (4.28) slightly to ease calculations. Since we are estimating the effect of the treatment on the treated, we can restrict our attention to the cases where $Z_i = 1$. Then (4.28) simplifies to

$$\mathbb{E}\left(Y \middle| \hat{Y}_c\right) = \tau_w + \hat{Y}_c. \tag{4.32}$$

The τ_w from this this model has the same value and interpretation as τ_w from (4.28). By default, the standard error associated with it will differ, but in either case we will use sandwich estimators to correctly compute it.

Following the derivation in Appendix C.2 and replacing Y_c with $\text{logit}^{-1}(X\beta_c)$, the second stage model has estimating equations

$$\psi(Y_i, \beta_c; \tau_w) = w_i(Y_i - \text{logit}^{-1}(X_i\beta_c) - \tau_w). \tag{4.33}$$

Following the derivations in Appendix B.1.2, we have that the bread are defined as

$$B_{11} = \mathbb{E} \sum_{i:Z_i=0} X_i X_i' \frac{\exp(X_i\beta_c)}{(1+\exp(X_i\beta_c))^2}, \tag{4.34}$$

$$B_{12} = 0, \tag{4.35}$$

$$B_{21} = \mathbb{E} \sum_{i:Z_i=1} w_i X_i' \frac{\exp(X_i\beta_c)}{(1+\exp(X_i\beta_c))^2}, \tag{4.36}$$

and

$$B_{22} = \mathbb{E} \sum_{i:Z_i=1} w_i = \frac{1}{2}. \tag{4.37}$$

To see why B_{22} simplifies, let n_s be the number of observations and n_{zs} be the number of treated members in strata s. We can rewrite δ_i as n_{zs}/n_s and $(1-n_{zs})/n_s$ (for observation i in strata s) when $Z_i = 1$ and 0 respectively. If we consider δ^{-1} and sum over all treated members, each strata will contribute n_{zs} identical additive components, so that $\sum_{i:Z_i=1} \delta^{-1} = \sum_s \frac{n_s}{n_{zs}} n_{zs} = n$. A similar calculation when summing over control members yields $\sum_{i:Z_i=0} \delta^{-1} = n$. Hence the sums of w over control and treatment are identical (and sum to 1 by definition).

The meat is

$$M_{11} = \sum_{i:Z_i=0} \text{Var}\left(Y_i - \text{logit}^{-1}(-X_i\beta_c)\right) X_i X_i' \tag{4.38}$$

which is estimated by

$$\hat{M}_{11} = \sum_{i:Z_i=0} \left(Y_i - \text{logit}^{-1}(-X_i\beta_c)\right)^2 X_i X_i', \tag{4.39}$$

and

$$M_{22} = \sum_{i:Z_i=1} \text{Var}\left(w_i \left(Y_i - \text{logit}^{-1}(X_i\beta_c) - \tau_w\right)\right) \tag{4.40}$$

estimated by

$$\hat{M}_{22} = \sum_{i:Z_i=1} w_i^2 \left(Y_i - \text{logit}^{-1}(X_i\beta_c) - \tau_w\right)^2. \tag{4.41}$$

To see why $\hat{M}_{22}$ drops the expectation squared, note that in weighted least squares, the expected value of weights times residuals is zero. We obtain a final estimate of

$$\hat{\sigma}^2_{\text{wls}} = 4\left(\hat{M}_{22} + \hat{B}_{21}\hat{B}_{11}^{-1}\hat{M}_{11}\hat{B}_{11}^{-T}\hat{B}_{21}^T\right). \tag{4.42}$$

4.4.4 Ignoring the Decision Criterion

There are situations where the choice of a linear or logistic model may be based upon desired properties of the treatment effect estimate rather than the decision criterion we describe in Section 4.4.2.

When the second stage is a linear model, the coefficient on treatment status Z is a consistent estimate of the average treatment effect, and if the second stage is weighted (as we recommend in Section 4.4.3), then it is a consistent estimate of the weighted average treatment effect.[5] This holds regardless of whether the treatment effects are linear on the probability scale. This same property does not hold for the logistic second stage model.

On the other hand, if the second stage is logistic, we can benefit from the re-

versibility of an odds ratio. If, for example, data are collected from a case-control study, disease rates given exposure cannot be estimated. However, since we do obtain estimates of exposure given disease rates, and the odds ratios for those two conditional odds are equivalent.[61, Ch. 2] A similar property does not exist for a linear second stage model.

4.5 Applied Example

We now re-examine the results of Gurm et al. [32]. In the paper, the authors are examining whether vascular closure devices (VCDs) can reduce the risk of vascular complications after arterial access. After matching those with VCDs and those without, the authors estimate the effect of the usage of VCDs on the existence of vascular complications by way of a conditional logistic regression model, conditioning on the matched sets. The results show a statistically significant reduction of the odds of a vascular complication, with an odds ratio of 0.78.

4.5.1 Detecting Treatment Effect on Linear Scale

The published results show that the effect of VCD usage is linear on the logit scale. The authors, not being aware of our recommendations in this work, do not ask whether the treatment effect might be better modeled by linear on the probability scale.

We first implement our recommendations in Section 4.4.2, comparing second-stage models based upon (4.26) and (4.25). Let Y be the binary response of a vascular complication, let X be an $n \times p$ matrix of covariates (such as a constant column for the intercept, prior congestive heart failure and the hospital in which the procedure was performed), and let Z be the treatment indicator, the use of a VCD. The first-stage model is

$$\text{logit}\left(\mathbb{E}(Y|X)\right) = X\beta. \tag{4.43}$$

The second stage models introduce λ from (4.22). Then, the new models are

$$\text{logit}\left(\mathbb{E}(Y|Z,\hat{Y}_c)\right) = Z\tau + \text{logit}(\hat{Y}_c), \tag{4.44}$$

$$\text{logit}\left(\mathbb{E}(Y|Z\lambda,\hat{Y}_c)\right) = (Z\lambda)\tau + \text{logit}(\hat{Y}_c). \tag{4.45}$$

We compare model fits, following our advice from Section 4.3.2 and using the estimated risk function based on logistic loss as the selection criteria. The results are shown in Table 4.1. Therefore there is evidence that the treatment effect may be better served by linearity on the probability scale.

	(4.44)	(4.45)
Estimated Risk	0.0917	0.0808

Table 4.1: Estimated risk based upon logistic loss for model (4.44) versus (4.45). The estimated risk is lower for the second model, suggesting that the treatment effect might be better modeled with a linear effect in probability.

4.5.2 Two-stage model

Now that we have evidence that the effect of treatment may be well fit as linear on the probability scale, we will use the suggestions of Section 4.4.3.

Continuing with the first stage model (4.43), we fit a linear second stage model,

$$\mathbb{E}\left(Y\middle|Z,\hat{Y}_c\right) = Z\tau + \hat{Y}_c, \tag{4.46}$$

computing the standard error as described in (4.42). We have that $\hat{\tau} = -0.002$ with standard error 0.0019. This is no longer significant. Further study could examine whether this approach lacks power compared to the conditional logistic model.

4.6 Conclusion

When dealing a binary response and studying treatment effect, typical analysis methods will force the treatment effect to be linear on the logit scale, or multiplicative on the probability scale. Using two-stage regression models, we introduce methodology to enable fitting the treatment effect linearly on the probability scale, while the relationship between response and other predictors remains on the logit scale. We showed that using the estimated risk based on logistic loss can yield a decision criteria to determine upon which scale the treatment effect is linear.

For stratified data, specifically matched sets, accounting for the stratification with binary response is typically handled with conditional logistic regression. We offer a two-stage alternative, which accounts for the strata via inverse probability weighting in the second stage. This two-stage approach enables an easier interpretation of interaction terms. Additionally, we gain the benefits discussed above, namely testing which scale the treatment effect is linear on. By using a sandwich approach to estimating the standard errors in the various second stage models, we open up the opportunity to expand the possible forms of both stage models.

APPENDICES

APPENDIX A

Appendix for Chapter II

A.1 Bias Correction

Although bias correction does not play a role in our method, we show here an attempt at bias correction, though we will ultimately show in Appendix A.2.1 that it does not improve coverage with the Wald confidence interval.

Consider again (2.38),

$$Y - X\beta_c = \tau + X\beta_c\eta + e. \tag{A.1}$$

The form of the estimate for η is not affected by the peculiarities of the PBPH method and thus the typical least squares parameter estimate suffices,

$$\hat{\eta} = \frac{\mathrm{Cov}(Y - X\hat{\beta}_c, X\hat{\beta}_c)}{\mathrm{Var}(X\hat{\beta}_c)}, \tag{A.2}$$

which is estimating the population η, defined by

$$\eta = \frac{\mathrm{Cov}(Y - X\beta_c, X\beta_c)}{\mathrm{Var}(X\beta_c)}. \tag{A.3}$$

Deriving the overall bias is quite difficult. Therefore we attempt only to minimize the bias. Specifically, we will bias correct the numerator and denominator of (A.2) separately, which leaves $\hat{\eta}$ biased (because the numerator and denominator are not independent), but reduces the overall bias.

Assume for simplicity and without loss of generality that X is centered overall, which allows us to further assume that the treatment group means of X converges in probability to 0. Thus we can claim that for any $\hat{\beta}$, $\sum_k \sum_i X_{ki}\hat{\beta}_k = 0$. Therefore, simple calculation shows us that, in the treatment group, the sample covariance (the numerator of (A.2)) can be expressed as

$$\frac{1}{n_t}\left(Y - X\hat{\beta}\right)'\left(X\hat{\beta}\right). \tag{A.4}$$

Since we are working solely in the treatment group, $Y = Y_t$, and trivially $Y = Y_t - Y_c + Y_c$, so that (A.4) becomes

$$\frac{1}{n_t}\left[\underbrace{(Y_t - Y_c)'(X\hat{\beta})}_{(*)} + (Y_c - X\hat{\beta})'(X\hat{\beta})\right]. \tag{A.5}$$

When we eventually take expectations, $(*)$ will contribute $(Y_t - Y_c)'(X\beta)$ by linearity and thus will not directly introduce any bias. There may be additional complications to the variance or for central limit theorem approximations, but we relegate that to further study.

Writing $X\hat{\beta}$ as $X(\beta - \beta + \hat{\beta}) = X\beta - X(\beta - \hat{\beta})$, we have that

$$\begin{aligned}(Y_c - X\hat{\beta})'(X\hat{\beta}) = \\ (Y_c - X\beta)'(X\beta) + \underbrace{(\beta - \hat{\beta})'X'X\beta - (Y_c - X\beta)'X(\beta - \hat{\beta})}_{(**)} - (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}).\end{aligned} \tag{A.6}$$

Again when we take expectations, $(**)$ will vanish since it is linear in $(\beta - \hat{\beta})$ and $\mathbb{E}(\beta - \hat{\beta}) = 0$. Therefore, taking expectations of both sides, we have

$$\frac{1}{n_t}\mathbb{E}\left((Y_c - X\hat{\beta})'(X\hat{\beta})\right) = \frac{1}{n_t}(Y_c - X\beta)'(X\beta) - \frac{1}{n_t}\mathbb{E}\left[(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]. \quad \text{(A.7)}$$

Thus, as an estimate for the covariance between $Y - X\beta$ and $X\beta$, the treatment-group covariance between $Y - X\hat{\beta}$ and $X\hat{\beta}$ is negatively biased with magnitude

$$\mathbb{E}\left[(\beta - \hat{\beta})'\Sigma_X(\beta - \hat{\beta})\right] \quad \text{(A.8)}$$

(where $\Sigma_X = \frac{X'X}{N_t}$ is the empirical covariance of the baseline covariates amongst the treatment group members), which is nothing more than the sum over i, j of all element-wise products of $\text{Cov}(\beta - \hat{\beta})$ and Σ_X from the treatment group. If we have an unbiased estimate of this, we will have an unbiased estimate of the magnitude of the bias.

Consider the denominator, which is the sample variance of $X\hat{\beta}$, with the centering assumptions above, can be written as

$$\frac{1}{n_t}\left(X\hat{\beta}\right)'\left(X\hat{\beta}\right). \quad \text{(A.9)}$$

Following the derivation of (A.6), we expand and drop terms which will vanish in expectation, leaving

$$(X\hat{\beta})'(X\hat{\beta}) = (X\beta)'(X\beta) - (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}), \quad \text{(A.10)}$$

so that

$$\frac{1}{n_t}\mathbb{E}\left((X\hat{\beta})'(X\hat{\beta})\right) = \frac{1}{n_t}(X\beta)'(X\beta) - \frac{1}{n_t}\mathbb{E}\left[(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]. \quad \text{(A.11)}$$

We are left with the same bias as in the numerator, yielding

$$\hat{\eta}^* = \frac{\text{Cov}(Y - X\hat{\beta}_c, X\hat{\beta}_c) + \hat{\mathbb{E}}\left[(\beta - \hat{\beta})'\Sigma_X(\beta - \hat{\beta})\right]}{\text{Var}(X\hat{\beta}_c) + \hat{\mathbb{E}}\left[(\beta - \hat{\beta})'\Sigma_X(\beta - \hat{\beta})\right]}, \tag{A.12}$$

as an estimator for η with less bias than $\hat{\eta}$, that is, $\mathbb{E}(\hat{\eta} - \eta) > \mathbb{E}(\hat{\eta}^* - \eta)$.

It would be convenient to be able to express the bias as a linear correction to $\hat{\eta}$. While in general there is no way to rewrite $\hat{\eta}^*$ as linear in $\hat{\eta}$, we can approximate it with a first order Taylor expansion, so that we have

$$\hat{\eta}^* \approx \hat{\eta} - \frac{\hat{\eta} - 1}{\text{Var}(X\hat{\beta}_c)}\hat{\mathbb{E}}\left[(\beta - \hat{\beta})'\Sigma_X(\beta - \hat{\beta})\right]. \tag{A.13}$$

When combined with the standard error correction, we ultimately have a method for obtaining an estimator for η which provides good coverage in the confidence interval setting.

A.1.1 A Simplifying Example

To consider a concrete example, let's consider $\hat{\beta}$ to come from a linear regression model between Y and X where $X \in \mathbb{R}_{n\times p}$. For notation, let Σ_{X_t} and Σ_{X_c} to be the empirical covariances of baseline covariates amongst treatment and control group members respectively. Then, we can simplify,

$$\text{Cov}(\beta - \hat{\beta}) = \text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1} = \sigma^2\frac{\Sigma_{X_c}^{-1}}{n_c - 1}. \tag{A.14}$$

To simplify notation (although likely not calculation), note that the element-wise product of two matrices is equivalent to the trace of their product. Assume $\hat{\sigma}^2$ is any unbiased estimator for σ^2, we therefore have that the bias existing in both the

numerator and denominator can be expressed as

$$\mathbb{E}(\beta - \hat{\beta})\Sigma_{X_t}(\beta - \hat{\beta}) = \hat{\sigma}^2 \frac{\text{tr}(\Sigma_{X_c}^{-1}\Sigma_{X_t})}{n_c - 1}. \tag{A.15}$$

Notice that this goes to 0 as $n_c \to \infty$ (provided of course that if $p \to \infty$, it does at a slower rate that n_c - not an unreasonable assumption in practice). Further, consider the trace term. If Σ_{X_t} is generally "larger" than Σ_{X_c} (rather than define "larger", just consider it in the hand-wavy sense of to have more extreme empirical covariances), then for a fixed σ^2, the bias will be higher, and when Σ_{X_t} is generally "smaller" than Σ_{X_c}, the bias will be lower. This follows intuition, namely that the bias grows as the treatment group becomes the dominant source of the sampling variability. We have less concern if the treatment group has lower sampling variability.

A.2 Failure of Wald-Style Confidence Intervals

We justify our claim that a Wald-style confidence interval is insufficient.

Generate a data set of size $n = 100$ using (2.25) and (2.27), for some value of $\eta \in (-1, 2)$. Perform the analysis using both uncorrected and corrected versions of the standard error, and check coverage of a Wald-type confidence interval using each version. (Note that a Wald-type uses the fully empirical estimator of the covariance, (2.17), as described in Section 2.4.2.) Repeat this 1,000 times for each choice of η, then repeat the entire procedure with $n = 1{,}000$ to check for sample size considerations. The resulting coverage percentages are plotted in Figure A.1.

The corrected standard error outperforms the uncorrected estimate, however coverage is still lacking.

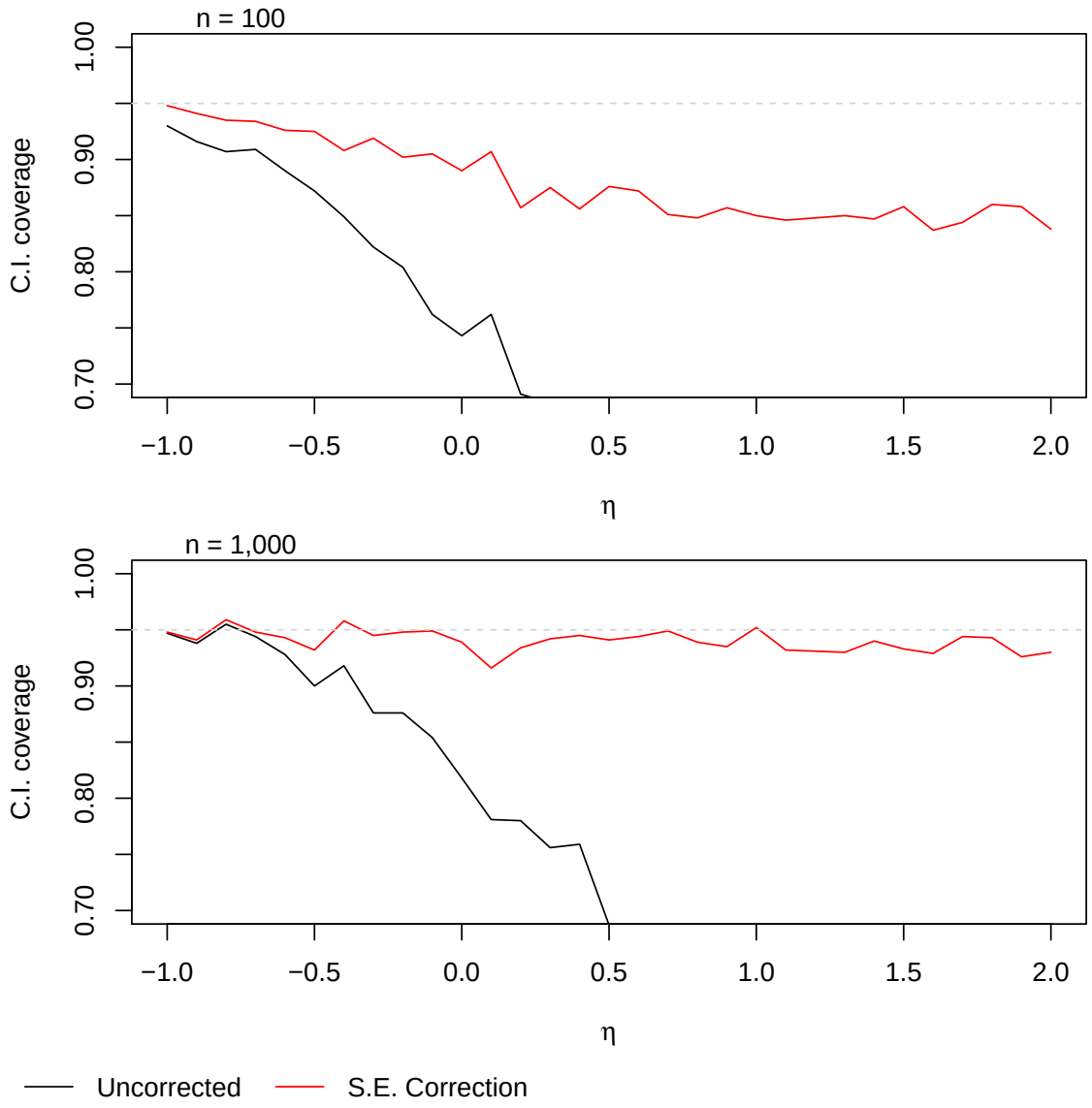

Figure A.1: Simulation results comparing coverage of confidence intervals built with the uncorrected and corrected standard error estimates, using samples sizes $n = 100$ and $n = 1000$.

A.2.1 Adding in bias correction

Adding the bias correction above, we still do not see proper Wald coverage. Results using the same simulation settings as above, we obtain the coverage percentages plotted in Figure A.2.

Once again, we have improved on the coverage over the standard error correction alone (barring the oddity of poor performance as η approaches -1, which is likely due to the unique properties of $\eta = -1$; see Section 2.5.1.1) we still do not have acceptable coverage.

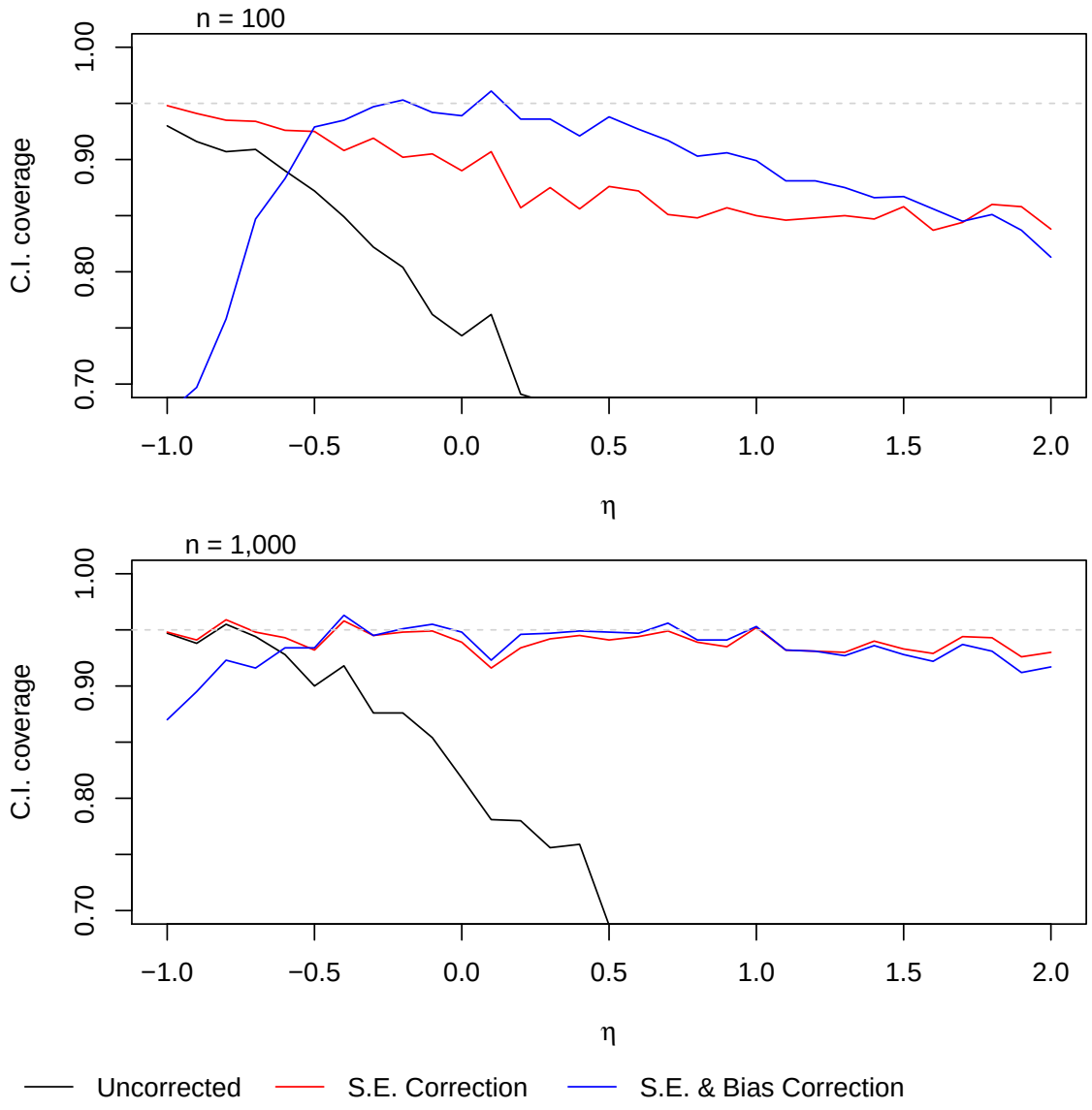

Figure A.2: Simulation results comparing coverage of confidence intervals built with the corrected standard error estimates, with and without bias correction, at the different sample sizes.

APPENDIX B

Appendix for Chapter III

B.1 Derivation of Sandwich Components with GLM First Stage

Assume that the responses Y share some distribution from the exponential family and

$$\mathbb{E}(Y_i) = \mu_i \tag{B.1}$$

$$g(\mu_i) = X\beta_c, \tag{B.2}$$

where $X \in \mathbb{R}_{n\times p}$ is the design matrix, including a first constant column for an intercept, and g is some monotone and twice-differential canonical link function. For example, if Y is logistically distributed, $g(t) = \log\left(\frac{t}{1-t}\right)$. If Y is Poisson then $g(t) = \log(t)$.

Let $h(t) = g^{-1}(t)$ to simplify notation. Then the second stage is now

$$Y - h(X\beta_c) = \tau + \eta h(X\beta_c). \tag{B.3}$$

To define the estimating equations, we return to first principles. The estimating

equation for the first stage model will be derived as the derivative of the log likelihood of $\beta_c|Y_i$. All members of the exponential family can have their distribution described as

$$f(y_i|\beta_c) = s(y_i)t(\beta_c)\exp\left[\sum_{k=1}^{K} a_k(y_i)b_k(\beta_c)\right] \tag{B.4}$$

$$= \exp\left[\left(\sum_{k=1}^{K} a_k(y_i)b_k(\beta_c)\right) + c(y_i) + d(\beta_c)\right], \tag{B.5}$$

where $c(y_i) = \log s(y_i)$ and $d(\beta_c) = \log t(\beta_c)$.[23]

The corresponding log likelihood is

$$l(\beta_c|y_i) = \left(\sum_{k=1}^{K} a_k(y_i)b_k(\beta_c)\right) + c(y_i) + d(\beta_c), \tag{B.6}$$

and the first stage estimating equation is the derivative with respect to β_c,

$$\phi(Y_i;\beta_c) = \frac{\partial}{\partial\beta_c} l(\beta_c|y_i) = \left(\sum_{k=1}^{K} a_k(y_i)\left(\frac{\partial}{\partial\beta_c} b_k(\beta_c)\right)\right) + \frac{\partial}{\partial\beta_c} d(\beta_c). \tag{B.7}$$

The second stage, remaining linear, is similar to that developed in Section 2.5.2.2,

$$\psi_i(Y_i,\beta_c;\tau,\eta) = \big(Y_i - h(X_i\beta_c) - \tau - \eta h(X_i\beta_c)\big)\begin{pmatrix} 1 \\ h(X_i\beta_c) \end{pmatrix}. \tag{B.8}$$

Estimators for the all parameters of interest, (β_c, τ, η), are solutions from

$$\begin{pmatrix} \\ \mathbf{0} \\ \\ \end{pmatrix} = \begin{pmatrix} \sum_{\{i:Z_i=0\}} \phi_i(Y_i;\beta_c) \\ \sum_{\{i:Z_i=1\}} \psi_i(Y_i,\beta_c;\tau,\eta) \end{pmatrix} = \begin{pmatrix} \Phi(Y;\beta_c) \\ \\ \Psi(Y,\beta_c;\tau,\eta) \end{pmatrix}. \tag{B.9}$$

As with the linear version, we approach this derivation using a blocked matrix.

The bread matrix has the form

$$B(\beta_c, \tau, \eta) = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}\frac{\partial}{\partial \beta_c}\Phi(Y; \beta_c) & \mathbb{E}\frac{\partial}{\partial (\tau,\eta)}\Phi(Y; \beta_c) \\ \mathbb{E}\frac{\partial}{\partial \beta_c}\Psi(Y, \beta_c; \tau, \eta) & \mathbb{E}\frac{\partial}{\partial (\tau,\eta)}\Psi(Y, \beta_c; \tau, \eta) \end{bmatrix}, \quad \text{(B.10)}$$

where $B_{11} \in \mathbb{R}_{p\times p}$, $B_{12} \in \mathbb{R}_{p\times 2}$, $B_{21} \in \mathbb{R}_{2\times p}$ and $B_{22} \in \mathbb{R}_{2\times 2}$. To simplify notation, the submatrices and their estimates of the bread and meat are written succinctly. For example, B_{11} is shorthand for $B_{11}(\beta_c, \tau, \eta)$ and $\hat{B}_{11}$ is shorthand for $B_{n_t,11}(\hat{\beta}_c, \hat{\tau}, \hat{\eta})$.

B_{11} involves only the first stage, and is

$$B_{11} = \mathbb{E} \sum_{\{i:Z_i=0\}} \left[\left(\sum_{k=1}^{K} a_k(y_i) \left(\frac{\partial^2}{\partial \beta_c^2} b_k(\beta_c) \right) \right) + \frac{\partial^2}{\partial \beta_c^2} d(\beta_c) \right]. \quad \text{(B.11)}$$

Since the first stage does not include (τ, η),

$$B_{12} = \mathbf{0}. \quad \text{(B.12)}$$

B_{21} is slightly more complicated, since β_c exists in both stages,

$$B_{21} = \mathbb{E} \sum_{\{i:Z_i=1\}} \begin{pmatrix} -(1+\eta)\dot{h}(X_i\beta_c) \\ \big(Y_i - \tau - 2(1+\eta)h(X_i\beta_c)\big)\dot{h}(X_i\beta_c) \end{pmatrix}. \quad \text{(B.13)}$$

Finally,

$$B_{22} = \mathbb{E} \sum_{\{i:Z_i=1\}} \begin{bmatrix} 1 & h(X_i\beta_c) \\ h(X_i\beta_c) & h(X_i\beta_c)^2 \end{bmatrix}. \quad \text{(B.14)}$$

The meat matrix $M^{(c)}(\beta_c, \tau, \eta)$ will be similarly blocked. The diagonal blocks, M_{11} and M_{22}, will be the variance of Φ and Ψ respectively. The off-diagonal blocks remain 0 as in the linear case, see Section 2.5.2.2.

M_{11}, being the variance of Φ, is simply

$$M_{11} = \text{Var}\left[\sum_{\{i:Z_i=0\}}\left(\sum_{k=1}^{K} a_k(y_i)\left(\frac{\partial}{\partial\beta_c} b_k(\beta_c)\right)\right) + \frac{\partial}{\partial\beta_c} d(\beta_c)\right]. \tag{B.15}$$

The bottom right piece involves all three parameters of interest

$$M_{22} = \text{Var}\left[\sum_{\{i:Z_i=1\}} \big(Y_i - h(X_i\beta_c) - \tau - \eta h(X_i\beta_c)\big)\begin{pmatrix} 1 \\ h(X_i\beta_c) \end{pmatrix}\right]. \tag{B.16}$$

Simplifying the meat without specifying the link function is quite difficult; we leave that task to after specifying a distribution for Y.

The covariance of (τ, η) is the lower right 2×2 sub-matrix of

$$B_{n_t}^{(c)}(\hat{\beta}_c, \hat{\tau}, \hat{\eta})^{-1} M_{n_t}^{(c)}(\hat{\beta}_c, \hat{\tau}, \hat{\eta}) B_{n_t}^{(c)}(\hat{\beta}_c, \hat{\tau}, \hat{\eta})^{-T}. \tag{B.17}$$

After some tedious but simple algebra, we arrive at

$$\text{Var}\left(\tau, \eta\right) = B_{22}^{-1}\left(M_{22} + B_{21}B_{11}^{-1}M_{11}B_{11}^{-T}B_{21}^{T}\right)B_{22}^{-T}. \tag{B.18}$$

B.1.1 Example: Ordinary Linear Model

When $Y|\beta_c$ is normal, the first stage model is the normal linear model. Therefore we can confirm the results in Chapter II. In this setting, g is the identity function (and similarly h), therefore, $Y|\beta_c$ has mean $X\beta_c$ and variance σ^2, though we consider σ^2 a nuisance parameter.

We have that

$$f(y_i|\beta_c) \propto \exp\left(-\frac{(y_i - x_i\beta_c)^2}{2\sigma^2}\right) \tag{B.19}$$

$$\propto \exp\left(-\frac{y_i^2}{2\sigma^2} + \frac{y_i x_i \beta_c}{\sigma^2} - \frac{(x_i\beta_c)^2}{\sigma^2}\right), \tag{B.20}$$

so that $k = 1$ and $a_1(y_i) = y_i$, $b_1(\beta_c) = \frac{x_i\beta_c}{\sigma^2}$, and $c(y_i) = -\frac{y_i^2}{2\sigma^2}$ and $d(\beta_c) = -\frac{(x_i\beta_c)^2}{\sigma^2}$.

The first stage estimating equation is therefore

$$\phi(Y_i; \beta_c) = Y_i\frac{X_i}{\sigma^2} - \frac{(X_i\beta_c)X_i}{\sigma^2} = (Y_i - X_i\beta_c)X_i. \tag{B.21}$$

The second equality holds due to the estimating equation equaling 0. The second stage is clearly

$$\psi_i(Y_i, \beta_c; \tau, \eta) = \left(Y_i - X_i\beta_c - \tau - \eta X_i\beta_c\right)\begin{pmatrix} 1 \\ X_i\beta_c \end{pmatrix}, \tag{B.22}$$

agreeing with the results in Section 2.5.2.2.

B.1.2 Example: Logistic Regression

Let $Y_i|\beta_C$ be distributed as a Bernoulli trial with success probability ρ_i where

$$\rho_i = \frac{1}{1 + \exp(-X_i\beta_c)}. \tag{B.23}$$

The link function g is logit, so that its inverse is

$$h(X_i\beta_c) = \frac{1}{1 + \exp(-X_i\beta_c)} = \text{logit}^{-1}(X_i\beta_c). \tag{B.24}$$

Therefore we have

$$f(y_i|\beta_c) = \rho_i^{y_i}(1 - \rho_i)^{1-y_i} \tag{B.25}$$

$$= \exp\left(y_i \log\left(\frac{\rho_i}{1 - \rho_i}\right) + \log(1 - \rho_i)\right), \tag{B.26}$$

with $k = 1$, $a_1(y_i) = y_i$, $b_1(\rho_i) = \log\left(\frac{\rho_i}{1-\rho_i}\right)$, $c(y_i) = 0$ and $d(\rho_i) = \log(1 - \rho_i)$. Substituting (B.23) into b_1 and d, we get that $b_1(\beta_c) = X_i\beta_c$ and $d(\beta_c) = \log\left(\frac{1}{1+\exp(X_i\beta)}\right)$.

The corresponding first stage estimating equation is

$$\phi(Y_i;\beta_c) = -Y_i X_i + \text{logit}^{-1}(X_i\beta_c)X_i \tag{B.27}$$

$$= \left(Y_i - \text{logit}^{-1}(X_i\beta_c)\right) X_i, \tag{B.28}$$

where the sign switch is due to $\phi(Y_i;\beta_c) = 0$, and the second stage estimating equation is

$$\psi_i(Y_i,\beta_c;\tau,\eta) = \left(Y_i - \text{logit}^{-1}(X_i\beta_c) - \tau - \eta\,\text{logit}^{-1}(-X_i\beta_c)\right) \begin{pmatrix} 1 \\ \text{logit}^{-1}(-X_i\beta_c) \end{pmatrix} \tag{B.29}$$

Looking at the bread and meat, we see some complications. First, B_{11} is no longer independent of β_c,

$$B_{11} = \mathbb{E} \sum_{\{i:Z_i=0\}} X_i X_i' \frac{\exp(X_i\beta_c)}{(1+\exp(X_i\beta_c))^2}. \tag{B.30}$$

Note that the fraction is scalar, while $X_i X_i'$ is $p \times p$.

For the off-diagonals, B_{12} is still 0, and

$$B_{21} = \mathbb{E} \sum_{\{i:Z_i=1\}} \begin{pmatrix} -(1+\eta)X_i \\ \left(Y_i - \tau - 2(1+\eta)\,\text{logit}^{-1}(-X_i\beta_c)\right) X_i \end{pmatrix} \frac{\exp(X_i\beta_c)}{(1+\exp(X_i\beta_c))^2}. \tag{B.31}$$

Both B_{11} and B_{21} have scaling terms of the same form, but are summed over the control and treatment groups respectively.

Finally, B_{22} is straightforward,

$$B_{22} = \mathbb{E} \sum_{\{i:Z_i=1\}} \begin{bmatrix} 1 & \text{logit}^{-1}(X_i\beta_c) \\ \text{logit}^{-1}(X_i\beta_c) & \left(\text{logit}^{-1}(X_i\beta_c)\right)^2 \end{bmatrix}. \tag{B.32}$$

Moving to the meat, we have that

$$M_{11} = \sum_{\{i:Z_i=0\}} \text{Var}\left(Y_i - \text{logit}^{-1}(-X_i\beta_c)\right) X_i X_i', \tag{B.33}$$

and

$$M_{22} = \sum_{\{i:Z_i=1\}} \text{Var}\left[\left(Y_i - \text{logit}^{-1}(-X_i\beta_c) - \tau - \eta\,\text{logit}^{-1}(-X_i\beta_c)\right)\begin{pmatrix} 1 \\ \text{logit}^{-1}(-X_i\beta_c) \end{pmatrix}\right]. \tag{B.34}$$

Since $Y_c = \text{logit}^{-1}(-X_i\beta_c)$, both pieces of the meat and B_{22} have forms that are similar to the linear case. However, the other two pieces of the bread have the additional multiplicative term, $\frac{\exp(X_i\beta_c)}{(1+\exp(X_i\beta_c))^2}$. This is simply the variance, so can be represented by $\rho_i(1-\rho_i)$ to ease computation.

B.1.3 Example: Poisson Regression

Next, let $Y_i|\beta_C$ be Poisson with expected value λ_i where

$$\lambda_i = e^{X_i\beta_c}. \tag{B.35}$$

The link function is a log, so its inverse is

$$h(\mu_i) = e^{X_i\beta_c}. \tag{B.36}$$

We have

$$f(y_i|\beta_c) = e^{Y_i X_i \beta_c} e^{-e^{i\beta_c}} Y!^{-1} \tag{B.37}$$

$$= \exp\left(Y_i X_i \beta_c - e^{X_i \beta_c} - \log Y!\right), \tag{B.38}$$

with $k = 1$, $a_1(Y_i) = Y_i$, $b_1(\beta_c) = X_i\beta_c$, $c(Y_i) = -\log Y!$ and $d(\beta_c) = -\exp(X_i\beta_c)$

The estimating equations are

$$\phi(Y_i; \beta_c) = Y_i X_i - X_i e^{X_i \beta_c} \tag{B.39}$$

$$= (Y_i - e^{X_i \beta_c}) X_i, \tag{B.40}$$

and

$$\psi_i(Y_i, \beta_c; \tau, \eta) = \left(Y_i - e^{X_i\beta_c} - \tau - \eta e^{X_i\beta_c}\right) \begin{pmatrix} 1 \\ e^{X_i\beta_c} \end{pmatrix}. \tag{B.41}$$

B_{11} is still no longer independent of β_c,

$$B_{11} = \mathbb{E} \sum_{\{i:Z_i=0\}} X_i X_i' e^{X_i\beta_c}. \tag{B.42}$$

B_{12} is still 0, and

$$B_{21} = \mathbb{E} \sum_{\{i:Z_i=1\}} \begin{pmatrix} -(1+\eta) X_i \\ \left(Y_i - \tau - 2(1+\eta) e^{X_i\beta_c})\right) X_i \end{pmatrix} e^{X_i\beta_c}, \tag{B.43}$$

and

$$B_{22} = \mathbb{E} \sum_{\{i:Z_i=1\}} \begin{bmatrix} 1 & e^{X_i\beta_c} \\ e^{X_i\beta_c} & (e^{X_i\beta_c})^2 \end{bmatrix}. \tag{B.44}$$

The meat diagonals are

$$M_{11} = \sum_{\{i:Z_i=0\}} \text{Var}\left(Y_i - e^{X_i\beta_c}\right) X_i X_i', \tag{B.45}$$

and

$$M_{22} = \sum_{\{i:Z_i=1\}} \text{Var}\left[\left(Y_i - e^{X_i\beta_c} - \tau - \eta e^{X_i\beta_c}\right)\begin{pmatrix} 1 \\ e^{X_i\beta_c} \end{pmatrix}\right]. \tag{B.46}$$

As with the logistic case, we have a result similar in form to the linear case, with an additional component on B_{11} and B_{12}, $e^{X_i\beta_c}$ which is the variance, λ_i.

B.2 Derivation of Sandwich Components for Clustered Data

B.2.1 Clustered Standard Errors

We extend the estimating equation and M-estimator framework into the clustered setting. Each M-estimator is the solution to an estimating equation, namely $\hat{\theta}$ is an M-estimator for θ if $\hat{\theta}$ solves

$$0 = \sum_{i=1}^{n} \phi_i(D_i, \theta), \tag{B.47}$$

where D_i are some independent data and ϕ_i are known functions. Now consider a set of n observations, where there are S clusters and n_s observations in cluster s. We can re-write the estimating equation (B.47) as

$$0 = \sum_{s=1}^{S} \left(\sum_{i=1}^{n_s} \phi_{si}(D_{si}; \theta)\right). \tag{B.48}$$

If we consider a least squares regression setting, where $Y_{si} = X_{si}\beta + \epsilon_{si}$, then we have that

$$\phi_{si}(Y_{si}, X_{si}; \beta) = (Y_{si} - X_{si}\beta)X_{si}. \tag{B.49}$$

The bread will be the derivative of this with respect to the parameter, so

$$B(\beta) = \sum_{s=1}^{S} \left(\sum_{i=1}^{n_s} X_{si} \right) = \sum_{i=1}^{n} X_i, \tag{B.50}$$

which is identical to the non-clustered version. Clustering has no effect on the bread. However, in the meat, we do see an effect as

$$M(\beta) = \sum_{s=1}^{S} \left(\sum_{i=1}^{n_s} (Y_{si} - X_{si}\beta) X_{si} \right)' \left(\sum_{i=1}^{n_s} (Y_{si} - X_{si}\beta) X_{si} \right). \tag{B.51}$$

Computationally, we are able to compute the meat easily by first summing the estimating equation over each cluster.

Finally, the above is asymptotically correct but often uses a finite sample adjustment. One often used adjustment is

$$\frac{S}{S-1} \cdot \frac{n-1}{n-p}, \tag{B.52}$$

where p is the number of parameters, including intercept. This should be equivalent to the rank of the design matrix, assuming the design matrix is of full rank (equivalently that we can obtain estimates for all coefficients).[17]

B.2.2 PBPH with Clustering

We can extend the PBPH method to allow clustering. As above, assume we have n observations, each belonging to one of S clusters, with n_s observations is cluster s. Let S_0 and S_1 represent the set of clusters which were randomly assigned to control and treatment respectively. Otherwise, notation remains identical to the non-clustered variation.

With these clusters, the stacked estimating equations to solve now become

$$\begin{pmatrix} \mathbf{0} \end{pmatrix} = \begin{pmatrix} \sum_{s=1}^{S_0} \left(\sum_{i=1}^{n_s} \phi_i(Y_i; \beta_c) \right) \\ \sum_{s=1}^{S_1} \left(\sum_{i=1}^{n_s} \psi_i(Y_i, \beta_c; \tau, \eta) \right) \end{pmatrix}, \tag{B.53}$$

where as before, we have

$$\phi_i(Y_i; \beta_c) = (Y_i - X_i'\beta_c)X_i, \tag{B.54}$$

$$\psi_i(Y_i, \beta_c; \tau, \eta) = (Y_i - X_i'\beta_c - \tau - \eta X_i'\beta_c)\begin{pmatrix} 1 \\ X_i'\beta_c \end{pmatrix}. \tag{B.55}$$

As mentioned in Appendix B.2.1, the bread matrix B will not be affected by this shift.

For the meat, M, we have that

$$M_{11} = \text{Var}\left(\sum_{s=1}^{S_0} \left(\sum_{i=1}^{n_s} \phi_i(Y_i; \beta_c) \right) \right) \tag{B.56}$$

$$= \sum_{s=1}^{S_0} \left(\text{Var} \sum_{i=1}^{n_s} \phi_i(Y_i; \beta_c) \right) \tag{B.57}$$

$$= \sum_{s=1}^{S_0} \left(\sum_{i=1}^{n_s} \phi_i(Y_i; \beta_c) \right)' \left(\sum_{i=1}^{n_s} \phi_i(Y_i; \beta_c) \right). \tag{B.58}$$

The equality of (B.56) to (B.57) is due to observations being independent across clusters. The final equality to (B.58) is due to the estimating equation having mean 0. A very similar form exists for M_{22},

$$M_{22} = \sum_{s=1}^{S_1} \left(\sum_{i=1}^{n_s} \psi_i(Y_i, \beta_c; \tau, \eta) \right)' \left(\sum_{i=1}^{n_s} \psi_i(Y_i, \beta_c; \tau, \eta) \right). \tag{B.59}$$

We tweak the finite sample adjustment in (3.4), yielding

$$\frac{S_0}{S_0 - 1} \cdot \frac{n_c - 1}{n_c - p}, \tag{B.60}$$

for M_{11} and

$$\frac{S_1}{S_1 - 1} \cdot \frac{n_t - 1}{n_t - 2}, \tag{B.61}$$

for M_{22}. Here, $n_t = \sum(1 - Z_i) = \sum_{s=1}^{S_0} n_s$ and $n_t = \sum Z_i = \sum_{s=1}^{S_1} n_s$. Recall that in the second stage model, $p = 2$, hence the denominator in the second term of (B.61). In terms of implementation, in the single stage version, the adjustment is multiplied to the final form of the covariance, $B^{-1}MB^{-T}$. However, in our two stage version, we can rewrite (B.18) with the scaling factors as as

$$\frac{S_1}{S_1 - 1} \cdot \frac{n_t - 1}{n_t - 2} B_{22}^{-1} M_{22} B_{22}^{-T} + \frac{S_0}{S_0 - 1} \cdot \frac{n_c - 1}{n_c - p} B_{22}^{-1} B_{21} B_{11}^{-1} M_{11} B_{11}^{-T} B_{21}^{T} B_{22}^{-T}. \tag{B.62}$$

APPENDIX C

Appendix for Chapter IV

C.1 Derivation of regression coefficients

C.1.1 Unweighted

Let $Y_i \in \mathbb{R}$ and $Z_i \in \{0, 1\}$ be the observed response and treatment status of individual i. Let Y_{ic} be the potential response of individual i under control.

The model of interest, without weights, is

$$\mathbb{E}(Y|Z, Y_c) = \beta Z + Y_c. \tag{C.1}$$

We have that

$$\hat{\beta} = \frac{\sum_i Z_i(Y_i - Y_{ic}) - n^{-1}\sum_i Z_i \sum_i (Y_i - Y_{ic})}{\sum_i Z_i^2 - n^{-1}\sum_i Z_i \sum_i Z_i}. \tag{C.2}$$

Now $Z_i^2 = Z_i$ and $\sum_i (Y_i - Y_{ic}) = \sum_i Z_i(Y_i - Y_{ic}) + \sum_i (1 - Z_i)(Y_i - Y_{ic})$. In the control group, the observed response is the potential response to control, so

$\sum_i (1 - Z_i)(Y_i - Y_{ic}) = 0$. Therefore,

$$= \frac{\sum_i Z(Y_i - Y_{ic)} - n^{-1} \sum_i Z_i \sum_i Z_i (Y_i - Y_{ic})}{\sum_i Z_i \left(1 - n^{-1} \sum_i Z_i\right)} \tag{C.3}$$

$$= \frac{\sum_i Z_i (Y_i - Y_{ic})(1 - n^{-1} \sum_i Z_i)}{\sum_i Z_i (1 - n^{-1} \sum_i Z_i)} \tag{C.4}$$

$$= \frac{\sum_i Z_i (Y_i - Y_{ic})}{\sum_i Z_i}. \tag{C.5}$$

$\hat{\beta}$ is the average of $Y - Y_c$ amongst the treatment group, or the estimated effect of the treatment on the treated.

C.1.2 Weights

Now, let w_i be the weight applied to individual i. Many of the same calculations and maneuvers carry over. The end result is that

$$\hat{\beta}_w = \frac{\sum_i w_i Z_i (Y_i - Y_{ic})}{\sum_i Z_i w_i}. \tag{C.6}$$

$\hat{\beta}_w$ is the weighted average of $Y - Y_c$ amongst the treated.

C.2 Estimating Equation for Weighted Least Squares

In OLS, we assume the variance is homoscedastic, that is, $\text{Var}_{\text{ols}}(\epsilon) = \sigma^2 \mathbf{I}$. Generalized least squares extends this to allow $\text{Var}_{\text{gls}}(\epsilon) = \Sigma$, with the only restrictions being that $\Sigma_{ii} > 0$ and $\Sigma_{ij} = \Sigma_{ji}$.[4] Weighted least squares is a special case of GLS where off-diagonals of Σ are 0, that is, $\text{Var}_{\text{wls}} = \vec{\sigma}^2 \mathbf{I}$ where $\vec{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_n)$.

Let $w_i = \sigma_i^{-2}$ so that log likelihood to minimize for weighted least squares can be rewritten as

$$l(\beta | y_i) \propto \sum_i w_i (y_i - x_i \beta)^2, \tag{C.7}$$

yielding estimating equations of

$$\phi(y_i; \beta) = w_i(y_i - x_i\beta)x_i \tag{C.8}$$

for observation i.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Alberto Abadie, Matthew M Chingos, and Martin R West. Endogenous stratification in randomized experiments. Technical report, National Bureau of Economic Research, 2013.

[2] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2002.

[3] Alan Agresti. Score and pseudo-score confidence intervals for categorical data analysis. *Statistics in Biopharmaceutical Research*, 3(2):163–172, 2011.

[4] Takeshi Amemiya. *Advanced econometrics*. Harvard university press, 1985.

[5] Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995.

[6] Simplice A Asongu. On the effect of foreign aid on corruption. *Economics Bulletin*, 32(3):2174–2180, 2012.

[7] Anamarie Auger, George Farkas, Margaret R Burchinal, Greg J Duncan, and Deborah Lowe Vandell. Preschool center care quality effects on academic achievement: An instrumental variables analysis. *Developmental psychology*, 50(12): 2559, 2014.

[8] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473): 138–156, 2006.

[9] Debabrata Basu. An essay on the logical foundations of survey sampling, part one. In *Selected Works of Debabrata Basu*, pages 167–206. Springer, 2011.

[10] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

[11] William A Belson. A Technique for Studying the Effects of a Television Broadcast. *Applied Statistics*, 5(3):195–202, 1956.

[12] Roger L Berger and Dennis D Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89 (427):1012–1016, 1994.

[13] A.S. Blinder. Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resources*, 8(4):436–455, 1973.

[14] Dennis D Boos. On generalized score tests. *The American Statistician*, 46(4): 327–333, 1992.

[15] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 2005.

[16] Stephen Burgess. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in medicine*, 32 (27):4726–4747, 2013.

[17] A Colin Cameron and Douglas L Miller. Robust inference with clustered data. Technical report, Working Papers, University of California, Department of Economics, 2010.

[18] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press, 2006.

[19] RJ Carroll, Suojin Wang, DG Simpson, AJ Stromberg, and D Ruppert. The sandwich (robust covariance matrix) estimator. *Unpublished manuscript*, 1998.

[20] W. G. Cochran. The Use of Covariance in Observational Studies. *Applied Statistics*, 18(3):270–275, 1969.

[21] William Cochran. Sampling techniques. *New York, Wiley and Sons*, 98:259–261, 1977.

[22] David Roxbee Cox and E Joyce Snell. *Analysis of binary data*, volume 32. CRC Press, 1989.

[23] Annette J Dobson and Adrian Barnett. *An Introduction to Generalized Linear Models*. CRC press, 2008.

[24] Susan Dynarski, Joshua M. Hyman, and Diane Schanzenbach. Experimental Evidence on the Effect of Childhood Investments on Post-secondary Attainment and Degree Completion. National Bureau of Economic Research Working Paper No. 17533, 2011.

[25] Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

[26] John Fox, Zhenghua Nie, and Jarrett Byrnes. *sem: Structural Equation Models*, 2016. URL `https://CRAN.R-project.org/package=sem`. R package version 3.1-7.

[27] Yoav Freund, Robert E Schapire, Yoram Singer, and Manfred K Warmuth. Using and combining predictors that specialize. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 334–343. ACM, 1997.

[28] Sally Galbraith, James A Daniel, and Bryce Vissel. A study of clustered data and approaches to its analysis. *The journal of Neuroscience*, 30(32):10601–10608, 2010.

[29] Xavier Giné, Jessica Goldberg, and Dean Yang. Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi. *American Economic Review*, 102(6):2923–2954, 2012.

[30] Sara Goldrick-Rab, Douglas N. Harris, James Benson, and Robert Kelchen. Conditional cash transfers and college persistence: Evidence from a randomized need-based grant program. Institute for Research on Poverty Discussion Paper No. 1393-11, 2011.

[31] Xu Guo, Wei Pan, John E Connett, Peter J Hannan, and Simone A French. Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Statistics in medicine*, 24(22):3479–3495, 2005.

[32] Hitinder S Gurm, Carrie Hosman, David Share, Mauro Moscucci, and Ben B Hansen. Comparative safety of vascular closure devices and manual closure among patients having percutaneous coronary intervention. *Annals of internal medicine*, 159(10):660–666, 2013.

[33] Frank E Harrell Jr. *rms: Regression Modeling Strategies*, 2016. URL `https://CRAN.R-project.org/package=rms`. R package version 4.5-0.

[34] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. 2001. *NY Springer*, 2001.

[35] Rodney A Hayward, David M Kent, Sandeep Vijan, and Timothy P Hofer. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Medical Research Methodology*, 6(18), 2006.

[36] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

[37] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[38] D Hosmer and S Lemeshow. *Applied Logistic Regression. New York, NY: A Wiley-Interscience Publication.* John Wiley & Sons Inc, 2000.

[39] Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233, Berkeley, Calif., 1967. University of California Press. URL `http://projecteuclid.org/euclid.bsmsp/1200512988`.

[40] Guido Imbens. Instrumental variables: An econometrician's perspective. Technical report, National Bureau of Economic Research, 2014.

[41] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[42] HK Iverson and RH Randles. The effects on convergence of substituting parameter estimates into u-statistics and other families of statistics. *Probability Theory and Related Fields*, 81(3):453–471, 1989.

[43] Robert W Keener. *Theoretical statistics: topics for a core course*. Springer, 2010.

[44] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of Economic Perspectives*, 15, 2004.

[45] Stephen W Lagakos. The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine*, 354(16):1667, 2006.

[46] Bruce G Lindsay and Annie Qu. Inference functions and quadratic score tests. *Statistical Science*, pages 394–410, 2003.

[47] Sharon Lohr. *Sampling: design and analysis*. Nelson Education, 2009.

[48] James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325, 1985.

[49] TK Mak and WK Li. A new method for estimating subgroup means under misclassification. *Biometrika*, 75(1):105–111, 1988.

[50] Kevin M Murphy and Robert H Topel. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1):88–97, 2002.

[51] R. Oaxaca. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14(3):693–709, 1973.

[52] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

[53] Charles C Peters. A Method of Matching Groups for Experiment With No Loss of Population. *The Journal of Educational Research*, 34(8):606–612, 1941.

[54] Stephen Portnoy. Asymptotic behavior of m-estimators of p regression parameters when p2/n is large. i. consistency. *The Annals of Statistics*, pages 1298–1309, 1984.

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

[56] R. Dennis Cook Richard R. Picard. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984. ISSN 01621459. URL http://www.jstor.org/stable/2288403.

[57] PR Rosenbaum. *Design of Observational Studies.(2010)*. New York Springer, 2010.

[58] Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

[59] Andrea Rotnitzky and Nicholas P Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77(3):485–497, 1990.

[60] D.B. Rubin. Estimating Causal Effects of Treatmetns in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66:688–701, 1975.

[61] James J Schlesselman. *Case-control studies: design, conduct, analysis*. Oxford University Press, 1982.

[62] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[63] Leonard A Stefanski and Dennis D Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

[64] James H Stock and Motohiro Yogo. Testing for weak instruments in linear iv regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 2005.

[65] Henry Theil. Economic forecasts and policy. 1958.

[66] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL http://CRAN.R-project.org/package=survival. version 2.38.

[67] Esteban Walker and Amy S Nowacki. Understanding equivalence and noninferiority testing. *Journal of general internal medicine*, 26(2):192–196, 2011.

[68] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[69] Philip Green Wright et al. Tariff on animal and vegetable oils. 1928.

[70] Achim Zeileis. Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(1):1–17, 2004. ISSN 1548-7660. doi: 10.18637/jss.v011.i10. URL https://www.jstatsoft.org/index.php/jss/article/view/v011i10.

[71] Achim Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(1):1–16, 2006. ISSN 1548-7660. doi: 10.18637/jss.v016.i09. URL https://www.jstatsoft.org/index.php/jss/article/view/v016i09.