Rethinking Nonlinear Instrumental Variables

by

Chunxiao Li

Department of Statistical Science
Duke University

Date: _____

Approved:

_____

David Dunson, Chair

_____

Cynthia Rudin, Supervisor

_____

Alex Volfovsky

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2019

ABSTRACT

Rethinking Nonlinear Instrumental Variables

by

Chunxiao Li

Department of Statistical Science
Duke University

Date: _____

Approved:

_____

David Dunson, Chair

_____

Cynthia Rudin, Supervisor

_____

Alex Volfovsky

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2019

Copyright © 2019 by Chunxiao Li
All rights reserved

Abstract

Instrumental variable (IV) models are widely used in the social and health sciences in situations where a researcher would like to measure a causal effect but cannot perform an experiment. Formally checking the assumptions of an IV model with a given dataset is impossible, leading many researchers to take as given a linear functional form and two stage least squares fitting procedure. In this paper, we propose a method for evaluating the validity of IV models using observed data and show that, in some cases, a more flexible nonlinear model can address violations of the IV conditions. We also develop a test that detects violations in the instrument that are present in the observed data. We introduce a new version of the validity check that is suitable for machine learning, and provide optimization-based techniques to answer these questions. We demonstrate the method using both the simulated data and a real-world dataset.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Instrumental variable (IV) analysis is a powerful methodology that allows estimation of causal effects in the presence of confounding variables. The IV framework assumes the presence of another variable, an "instrument" that varies in a way that is unrelated to the outcome of interest, except that it influences who receives treatment. One example of an IV is the so-called encouragement design, which is common in public health. A researcher is interested in the impact of a treatment (say taking a flu vaccine) on the likelihood of getting the flu. A regression that predicts likelihood of the flu based on whether or not a person got the vaccine will not necessarily reveal the true causal effect. This is because other – unmeasured – variables also impact the likelihood of getting the flu and are correlated with getting a vaccine (e.g. people who overall pay more attention to their health could be more likely to get the vaccine, but also take other measures to prevent the flu). In the encouragement design, a randomly selected set of individuals receive a reminder or other prompting to get a flu shot. The encouragement produces variation in the probability that some people will get the flu shot, but since the encouragement is assigned randomly it cannot be correlated with any other confounding variables. The variation in probability to get the flu shot that comes from being assigned the encouragement can then be used to identify the causal effect of getting a flu shot on getting the flu. Other common instruments include a policy change (e.g. a tax) that creates a change in a behavior that is the treatment of interest. Rainfall is commonly used as an instrument for changes in agriculture income. The instrumental variable framework allows us to analyze the treatment effect as seen through the lens of the instrument.

Two critical assumptions are required for IV models. First, we assume that the association between the instrument and the treatment variable is nontrivial. That is, that the (exogenous) variation in the instrument leads to meaningful variation in the treatment

variable, which means the variation is sufficiently strong so as to not be caused by noise. This is known as the relevance assumption. The relevance assumption can be assessed using the observed instrument and treatment directly. Second, we must assume that the only source of variation in the outcome from the instrument is through changes in the treatment variable. This is known as the exclusion restriction. In practice, the exclusion restriction cannot be verified with the data that a researcher has at hand. It is inherently a statement about unobserved variables (if a researcher has access to a known confounder, after all, she could simply include it). In this paper, we describe a statistical framework that, while still not able to completely verify the exclusion restriction, can provide the researcher empirical evidence about the quality of the instrument given the data at hand.

The most common way to analyze IV data is using two stage least squares. The first stage predicts treatment, based on the instrument and measured covariates. The second stage then models the outcomes as a linear function of covariates and the predicted values of the treatment from the first stage. Under the traditional least squares settings, both of the two critical assumptions use the correlation as the measurement. The relevance assumption states that there exists a strong correlation between the instrument and the treatment. The exclusion restriction states that the instrument is not directly correlated to the outcomes, in other word, the instrument is not correlated to the error term in the second stage. However, these assumptions can be problematic in nonlinear instrumental variable models. First, the assumptions themselves do not make sense for nonlinear models because linear correlation may not be relevant for nonlinear models. Second, it is possible for these assumptions to indicate that a valid instrument is actually invalid by construction. When that happens, there is no easy way to fix the problem. This means that in practice, many researchers simply do not check these assumptions and impose them intuitively instead, possibly leading to the discard of good instruments or the use of poor instruments. This leads naturally to important questions for IV analysis: What is the right validity check for nonlinear instrumental variable analysis? Can we fix our estimates so that good instruments pass the validity check? An instrument should not appear to be broken when we know it

is actually valid. Perhaps there is a flaw in our analytical procedures that can be fixed. Conversely, can we check whether an instrument is bad, in that a reasonable analysis would never find that it passes the validity check?

In Chapter 2, we generalize these assumptions for validity of the instrument to work for general, nonlinear models. Correlation, which is used in traditional IV formulations, is a measure of linear agreement. For example, in the exclusion restriction, if the instrument and error terms are correlated, it does not mean that their values are close together; correlation is not a good distance measure for nonlinear models. In order for us to create more complicated models, we need a measure of how similar the instrument is to the error terms. We use *prediction validity* for this task – if an instrument can predict the error terms, the instrument is defined to be invalid. In particular, if the instrument can predict the error terms approximately as well as the function 0 (that is identically 0), then the instrument is considered valid; it cannot predict the error terms any better than the function 0 can. Because we use prediction error to check validity of the instrument, we are not restricted to linear models; if an instrument appears to be bad due to the poor choice of a linear modeling procedure, nonlinear models can be used for both stages instead.

In Chapter 3, we introduce the definitions and assumptions which holds for the remainder of this paper. In Chapter 4, we propose a new version of the two-stage method for the more general non-linear framework, using prediction validity to check for valid instruments. This new two-stage method incorporate the prediction validity check as a constraint into the objective function to ensure that the instrument appears to be valid.

Also in Chapter 4, we present a one-stage procedure for IV analysis. For IV, we require that the instrument appears to be valid, and we also require high quality predictions for the treatment variable. These goals can by incorporated into a single mathematical program, using one constraint on prediction validity to ensure that the instrument appears to be valid and another to ensure that the instrument has a strong first stage.

In Chapter 5 and Chapter 6, we present the Lemmas and Theorems about the new version of the two-stage IV method and illustrate its limitations under some model con-

structions.

In Chapter 7, we provide the pseudo codes for the new versions of the two-stage and one-stage IV methods and kernelization options if desired.

In Chapter 8 and Chapter 9, we apply our two-stage and one-stage methods on some simulated datasets and a real-world dataset. On the one hand, we show that our methods with more complicated model constructions do often outperform the traditional two-stage method in terms of prediction power. On the other hand, we also discuss the limitations of our methods.

In Chapter 10, we provide a summary of main ideas in this paper and point out directions for the future work.

Chapter 2

Validity Check

In this chapter, we first introduced the framework of the traditional two stage least squares method and the original versions of two critical assumptions which use the correlation as the measurement. Then we generalize the original assumptions to the non-linear framework. The new versions of the two critical assumptions use loss functions to measure the predictability, which provides a more general measurement of the relationship between two variables. Finally, we define an empirical validity check from the new version exclusion restriction.

The notations used in the remainder of this paper are illustrated as below.

1. Notations of Sample Spaces: $\mathcal{X}$ is the feature space, $\mathcal{Z}$ is the space of possible values for the instrument, $\mathcal{T}$ is the space of possible values for the treatment, and $\mathcal{Y}$ are possible outcomes.

2. Notations of Populations: Random variables are capitalized, whereas realizations are lower cases. For example, the covariates X is a random variable whose domain is $\mathcal{X}$, whereas x is a realization of the covariates X. $\{x, z, t, y\}$ are realizations of random variables $\{X, Z, T, Y\}$ respectively.

3. Notations of Samples: Samples are represented by matrices or vectors, whereas the i-th observation in a sample is a lower case with subscript i. For example, $\vec{X}$ is a sample of the covariates X, whereas x_i is the i-th observation in the sample. $\{x_i, z_i, t_i, y_i\}$ are the i-th observation of samples $\{\vec{X}, \vec{Z}, \vec{t}, \vec{y}\}$ respectively.

4. Notations of Distributions: $\mathcal{D}_{X,Z}$ is a joint distribution over which X and Z are drawn. $\mathcal{W}_1$ and $\mathcal{W}_2$ are distributions of white noises that have zero mean, zero covariance and finite variance.

5. Notations of Models: The population models for T, Y, and U are represented by

$f_{\omega_{\text{true}}}$, $g_{\beta_{\text{true}}}$, and $h_{\alpha_{\text{true}}}$ respectively. The sample models for $\vec{t}$, $\vec{y}$, and $\vec{u}$ are represented by f_ω, g_β, and h_α respectively.

2.1 The Traditional Two Stage Method

We use a running example of tax laws as an instrumental variable z, smoking as a treatment t, and health conditions as outcomes y. Tax laws affect smoking (or cigarette consumption) through the price of cigarettes. Smoking has an influence on health conditions. Tax laws do not affect health conditions directly, they influence only the amount of smoking, which influences health conditions, given known covariates x. The data available are $\{x_i, z_i, t_i, y_i\}_{i=1}^n$ containing the known covariates, values of the instrument, the treatment, and outcomes, respectively, for each individual. For now, we assume there are no unknown covariates x_{UNi} outside the dataset that would correlate the instrument with outcomes directly.

In the context of instrumental variables, a popular form of estimation is known as the two-stage least squares (2SLS) regression. In typical two-stage least squares regression, we would build two linear models in Stage One and Stage Two respectively and solve them using the method of least squares. In Stage One, we would build a linear model to predict the amount of smoking from tax law information and covariate information, $\hat{t} = f_{\hat{\omega}}(x, z) = \langle \hat{\omega}, [x, z] \rangle$, where z is the value of the instrument (the presence of higher taxes), x are covariates (gender, age, etc.), and $\hat{\omega}$ is the estimated coefficient for the linear model. In Stage Two, we would estimate health conditions as a linear function of predicted amount of smoking and covariate information, $\hat{y} = g_{\hat{\beta}}(x, \hat{t}) = \langle \hat{\beta}, [x, \hat{t}] \rangle$, where $\hat{t}$ is the predicted values of the treatment, x are covariates, and $\hat{\beta}$ is the estimated coefficient of the linear model. This would provide an estimate for the effect of smoking on health conditions through the lens of the instrument.

As shown above, the fitted models of the traditional two-stage method are:

$$\hat{t}(x, z) = f_{\hat{\omega}}(x, z)$$
$$\hat{y}(x, \hat{t}) = g_{\hat{\beta}}(x, \hat{t}).$$

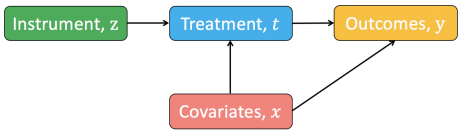

Figure 2.1: Graph without Unknown Covariates

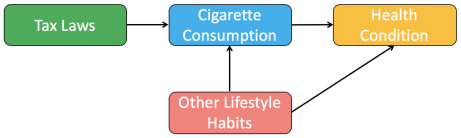

Figure 2.2: Example without Unknown Covariates

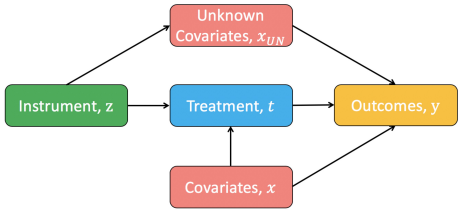

Figure 2.3: Graph with Unknown Covariates

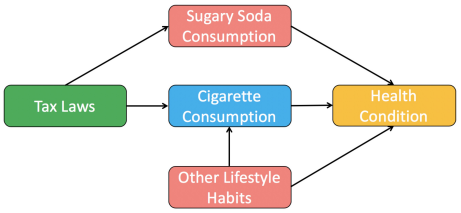

Figure 2.4: Example with Unknown Covariates

Let us consider the data as having been generated from the true models below:

$$(X, Z) \sim \mathcal{D}_{X,Z}$$

$$U \sim \mathcal{W}_1, V \sim \mathcal{W}_2$$

$$T = f_{\omega_{\text{true}}}(X, Z) + V$$

$$Y = g_{\beta_{\text{true}}}(X, T) + U$$

where ω_{true} and β_{true} are true coefficients of the linear models in Stage One and Stage Two respectively, and U and V are random error terms.

Now let us consider the potential interference or influence of outside unknown covariates X_{UN}. In the tax laws example, if the same tax laws also affect sugary soda consumption through the price of sugary soda, and drinking sugary soda has an influence on health conditions, then the tax laws can also affect health conditions through drinking sugary

soda. As a result, if tax laws do affect health conditions, we cannot determine whether the direct reason is by reducing cigarette consumption or sugary soda consumption. Let us consider the data as having been generated from the true models that contain the unknown covariates X_{UN} as follows:

$$
\begin{aligned}
&(X, X_{UN}, Z) \sim \mathcal{D}_{X,X_{UN},Z} \\
&U_{WN} \sim \mathcal{W}_1, V \sim \mathcal{W}_2 \\
&T = f_{\omega_{\text{true}}}(X, Z) + V \\
&Y = g_{\beta_{\text{true}}}(X, T) + U(X_{UN}(Z)) \\
&U(X_{UN}(Z)) = h(X_{UN}(Z)) + U_{WN}
\end{aligned}
$$

where V and U are error terms, V is random error term, U consists of a function of unknown covariates X_{UN}, $h(X_{UN}(Z))$ and random error terms U_{WN}, and X_{UN} is a function of the instrument Z. Thus the error term U can also be written as a function the instrument Z. Here the unknown covariates do not influence treatment, but they could.

2.2 Two Assumptions

Besides the (strong) assumption of the linear model form, it is well-known that there are two critical assumptions for using the traditional two-stage least squares regression: the relevance assumption and the exclusion restriction. Both the population versions and the sample versions of the assumptions are given as below:

Population Version of Relevance Assumption The instrument Z must be correlated with the treatment T in the first stage, conditionally on the other covariates x, i.e., for all x, $Cov(Z, T \mid x) = E((Z - E(Z))(T - E(T)) \mid x) \neq 0$. Thus, the instrument is *relevant.*

Population Version of Exclusion Restriction The instrument Z must not be correlated with the error term U in the second stage, $Cov(Z, U) = E((Z - E(Z))(U - E(U))) = 0$, where the error term U can be either random error terms or a function of

unknown covariates X_{UN}. The unknown covariates X_{UN} are possibly a function of the instrument Z.

Sample Version of Relevance Assumption The instrument $\vec{Z}$ cannot have all zero coefficients in the first stage i.e. For the model $\vec{t} = \omega_1 \vec{X} + \omega_2 \vec{Z} + \vec{v}$ in the first stage, at least one of the element in the vector ω_2 is not zero. The sample version of relevance assumption is assessed by the F-test with null hypothesis $H_0 : \omega_2 = \vec{0}$.

Sample Version of Exclusion Restriction The instrument $\vec{Z}$ must not be correlated with the error term $\vec{u}$ in the second stage, i.e. $Cov(\vec{Z}, \vec{u}) = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})(u_i - \bar{u}) = 0$, where the error term $\vec{u}$ can be either realizations of a random error term or a function of unknown covariates $\vec{X}_{UN}$. The unknown covariates $\vec{X}_{UN}$ are possibly a function of the instrument $\vec{Z}$.

The exclusion restriction above implies that the only way the instrument Z affects outcomes Y is through the treatment T but not any other unknown covariates X_{UN}. If there exist other unknown covariates X_{UN} through which the instrument Z also has an influence on outcomes Y, then the exclusion restriction does not hold. In the example above, the existence of sugary soda, which is an unknown covariate that is a function of tax laws and affects outcomes, would violate the exclusion restriction. Recall the notation above that the error term U is a function of the instrument Z, then $Cov(Z, U) = cov(Z, U(X_{UN}(Z))) \neq 0$. However, if the unknown covariates X_{UN} become known, we can control for them as usual, and the exclusion restriction will still hold, conditioned on X_{UN}, i.e., $Cov(Z, U \mid X_{UN}) = Cov(Z, U(X_{UN}) \mid X_{UN}) = 0$.

2.3 Non-linear Version

Both assumptions for the traditional two-stage method rely on the linear framework, using linear correlation between variables. The relevance assumption tests whether the correlation between the instrument Z and the treatment T equals 0, while the exclusion restriction tests whether the correlation between the instrument Z and the error term U equals 0.

To generalize the two-stage method to handle non-linear functions, we need to change the basic framework: correlation measures linear agreement, whereas we would like to consider nonlinear agreement. It is possible that correlation between two variables is weak, but that there is a strong nonlinear dependence between them. For example, if Y is a quadratic function of X, then the correlation between Y and X can be weak. We would like to develop modeling approaches that can handle this nonlinear dependence. Thus, we would like to create additional versions of both assumptions that can consider possible complicated nonlinear relationships for variables. This will be useful for ensuring that our models did not miss some transformation of the original covariates that is useful. The new nonlinear relevance assumption can also be assessed directly, while the new nonlinear exclusion relation has an empirical counterpart and can be checked.

In our new framework, we will no longer use (linear) correlation, we will use something more general, which is prediction loss. To determine how well several variables predict another variable, we consider how well they, together, can be used to predict it using functions of these variables from a pre-specified class. As in linear regression, we restrict the class of models to prevent overfitting, though overfitting can be restricted through other means as well (e.g., cross-validation).

Both the new population versions and the new sample versions of the assumptions are given as below:

Machine Learning Population Version of Relevance Assumption This assumption concerns the relationship between Z and T. It states that T can be predicted fairly well by X and Z.

$$\min_{f \in \mathcal{F}} loss(T, f(X, Z)) \leq \epsilon,$$

where $\mathcal{F}$ is a class containing all possible models and $loss$ is a real-valued loss function. Here, ϵ indicates a positive threshold.

Written another way (that will be generalized later), the assumption is:

$$loss(T, f_{\omega_{\text{true}}}(X, Z)) \leq \epsilon$$
$$\text{where } f_{\omega_{\text{true}}} \in \arg\min_{f \in \mathcal{F}} loss(T, f(X, Z)).$$

Machine Learning Population Version of Exclusion Restriction This assumption concerns the relationship between U and Z. It states that the instrument Z and known covariates X cannot be used to predict the error term U any better than a model that is identically 0 can, with a predetermined tolerance ϵ'.

$$loss(U, h_{\alpha_{\text{true}}}(X, Z)) \geq loss(U, 0) - \epsilon'$$
$$\text{where } U = Y - g_{\beta_{\text{true}}}(X, T)$$
$$\text{where } g_{\beta_{\text{true}}} \in \arg\min_{g \in \mathcal{G}} loss(Y, g(X, T))$$
$$\text{and } h_{\alpha_{\text{true}}} \in \arg\min_{h \in \mathcal{H}} loss(U, h(X, Z))$$

where $\mathcal{G}$ and $\mathcal{H}$ are classes containing all possible models and $loss$ is a real-valued loss function. Here, ϵ' indicates a positive threshold. This assumes we do not know the data generation process, but this does not matter because $\mathcal{G}$ and $\mathcal{H}$ include all possible functions, including the true functions from the data generation process $f_{\omega_{\text{true}}}$ and $g_{\beta_{\text{true}}}$.

Machine Learning Sample Version of Relevance Assumption This assumption concerns the relationship between $\vec{Z}$ and $\vec{t}$. It states that $\vec{t}$ can be predicted fairly well by $\vec{X}$ and $\vec{Z}$.

$$\min_{f \in F} loss(\vec{t}, f(\vec{X}, \vec{Z})) \leq \epsilon$$

or equivalently,

$$loss(\vec{t}, f_{\omega}(\vec{X}, \vec{Z})) \leq \epsilon$$
$$\text{where } f_{\omega} \in \arg\min_{f \in F} loss(\vec{t}, f(\vec{X}, \vec{Z}))$$

where F is a flexible class of models (where the usual reasonable measures have been taken to prevent overfitting) and $loss$ is a real-valued loss function. Here, ϵ indicates a positive threshold.

Machine Learning Sample Version of Exclusion Restriction This assumption concerns the relationship between $\vec{u}$ and $\vec{Z}$. It states that the instrument $\vec{Z}$ and known covariates $\vec{X}$ cannot be used to predict the error term $\vec{u}$ any better than a model that is identically 0 can, with a predetermined tolerance ϵ'.

$$
\begin{aligned}
&loss(\vec{u}, h_\alpha(\vec{X}, \vec{Z})) \geq loss(\vec{u}, 0) - \epsilon' \\
&\text{where } \vec{u} = \vec{y} - g_\beta(\vec{X}, \vec{t}) \\
&\text{where } g_\beta \in \text{argmin}_{g \in G} loss(\vec{y}, g(\vec{X}, \vec{t})) \\
&\text{and } h_\alpha \in \text{argmin}_{h \in H} loss(\vec{u}, h(\vec{X}, \vec{Z}))
\end{aligned}
$$

where G and H are flexible classes of models (where the usual reasonable measures have been taken to prevent overfitting) and *loss* is a real-valued loss function. Here, ϵ' indicates a positive threshold.

The new version of the exclusion restriction states that no matter how hard we try to minimize the loss using the instrument Z and known covariates X, we still cannot achieve a loss lower than what we can achieve using the model that is identically 0. This assumption is true if the error term U is random. However, if there exist unknown covariates X_{UN} through which the instrument Z also has an influence on outcomes Y, then this exclusion restriction does not hold any more. Recall the notation above that the error term U is a function of the instrument Z through the unknown covairates X_{UN} and can be written as $U(X_{UN}(Z)) = h(X_{UN}(Z)) + U_{WN}$. Then the error term U can be predicted by the model $h(X_{UN}(Z))$ of the instrument Z. Let $h_{\alpha_{\text{true}}}(X, Z) = h(X_{UN}(Z))$, then $loss(U, h_{\alpha_{\text{true}}}(X, Z)) = loss(U - h_{\alpha_{\text{true}}}(X, Z), 0) = loss(U - h(X_{UN}(Z)), 0) = loss(U_{WN}, 0)$ which is significantly smaller than $loss(U, 0)$. Thus, the exclusion restriction is violated. However, if we control for the unknown covariates X_{UN} by adding them as known covariates, then the model in second stage will be $g'_{\beta'}(X, X_{UN}, T) = g_{\beta_{\text{true}}}(X, T) + h(X_{UN})$ and in that case, the remainder U_{WN} will be random error terms and the exclusion restriction will still hold.

Considering unknown covariates X'_{UN} that can not be predicted by the instrument Z but can help predict outcomes Y, then the error term U is a function of the unknown

covariates X'_{UN} but not a function of Z. Therefore, the exclusion restriction is still valid.

2.4 Validity Check

Although both the population versions and the sample versions of the assumptions are given in previous sections, only the sample versions can be assessed with the available data. In practice, with both of the instrument $\vec{Z}$ and the treatment $\vec{t}$ observable, the sample version of the relevance assumption can be assessed directly. However, due to the fact that the true error term $\vec{u}$ is never observable, the sample version of the exclusion restriction is not testable. Recall the formula of the true error term $\vec{u} = \vec{y} - g_\beta(\vec{X}, \vec{t})$, where $g_\beta(\vec{X}, \vec{t})$ is the true model in the second stage. In practice, we would use the predicted value of the treatment $\hat{t}$ by the instrument instead of the true treatment $\vec{t}$ in the second stage, which represents the influence of the instrument through the treatment on outcomes. Then we would use the correspondingly estimated error term $\hat{u} = \vec{y} - g_\beta(\vec{X}, \hat{t})$ than the true error term $\vec{u}$. Here, we use the remainder r to represent the estimated error term $\hat{u}$ and its estimation $\hat{r}$ to represent the function $h_\alpha(\vec{X}, \vec{Z})$. Therefore, in order to test the sample version of the exclusion restriction, we use the following empirical validity check.

Machine Learning Empirical Validity Check

$$loss(\vec{r}, \hat{r}) \geq loss(\vec{r}, 0) - \epsilon' \quad \text{(predict remainders no better than null model)}$$

$$\text{where } \vec{r} = \vec{y} - \hat{y}, \;\; \hat{y} = g_\beta(\vec{X}, \hat{t}) \text{ and } \hat{r} = h_\alpha(\vec{X}, \vec{Z}) \quad \text{(remainder)}$$

$$\text{where } g_\beta \in \text{argmin}_{g \in G} loss(\vec{y}, g(\vec{X}, \hat{t})) \quad \text{(modeled outcomes)}$$

$$\text{and } h_\alpha \in \text{argmin}_{h \in H} loss(\vec{r}, h(\vec{X}, \vec{Z})) \quad \text{(modeled remainders)}$$

where G and H are flexible classes of models (where the usual reasonable measures have been taken to prevent overfitting). Here, ϵ' indicates a positive threshold.

Now, both the machine learning sample versions of the relevance assumption and the empirical validity check can be assessed. However, if we evaluate the relevance assumption and the empirical validity check after fitting the models, the fitted models ($f_{\hat{\omega}}$ and $g_{\hat{\beta}}$)

will be used instead of the true models (f_ω and g_β), which results in extra computational error when estimating $\hat{\omega}$ and $\hat{\beta}$. Therefore, in order to test the true model coefficients, we consider to incorporate them as constraints in the first and second stages respectively in the methodology chapter.

2.4.1 Connection to Adversarial Machine Learning

Our general machine learning validity connects to adversarial learning, in that in order for the ML empirical validity check to be valid, the remainders from the second stage must have been *generated* in a way that we cannot use them to *discriminate* between the outcomes any better than a model that is identically zero.

To turn this into an adversarial min/max formulation, one would maximize the loss for $loss(\vec{r}, h_\alpha(\vec{X}, \vec{Z}))$ with respect to r. The discriminator, which consists of the optimization problem for h_α, would aim to predict the remainders r. If the generator wins, then it is not possible for us to predict r any better than 0 can. If the discriminator wins, then h_α can approximate r and the validity check fails.

The validity check is a feasibility condition, not an optimality condition. This is why the generator's "max" does not appear, instead replaced by an inequality (in the first line of the validity check).

Chapter 3

Definitions and Assumptions

Definition 1. *A general additive model (GAM) is a general linear model that can be written as a linear combination of both linear and non-linear features i.e. A general additive model with input variable X and target variable y has the following form:*

$$\hat{y} = b_0 + b_1 \text{feature}_1(X) + \cdots + b_q \text{feature}_q(X)$$

where $\text{feature}_j(X)$*,* $j = 1, \cdots, q$ *is a linear or non-linear function of* X *and* $X = (x_1, \cdots, x_p)$*.*

Note that the general additive model (GAM) defined above is different form the generalized additive model with the following form:

$$\hat{y} = b_0 + b_1 \text{feature}_1(x_1) + \cdots + b_p \text{feature}_p(x_p)$$

where $\text{feature}_j(x_j)$*,* $j = 1, \cdots, p$ *is a linear or non-linear function of* x_j*.*

Assumption 1. *There is no multi-collinearity (or perfect collinearity) between input variables. This assumption holds for the remainder of this paper.*

Assumption 2. *The square matrix of the input variables is invertible (non-singular). This assumption holds for the remainder of this paper.*

Chapter 4

Methodology

4.1 Two Stage Method

We will first introduce the new two stage formulation, where the estimates for $\hat{t}$ and $\hat{y}$ are based on general loss minimization. In the general formulation, both stages can use nonlinear models. If using linear or general additive models (GAMs) in both stages, with the squared loss, the computations simplify and we can gain more insight. The empirical validity check is used as a constraint.

4.1.1 General ML Two-Stage Least Squares

Stage One

The optimization problem can be written as follows:

$$\omega \in \arg\min_{\omega} \sum_i loss(t_i, \hat{t}_i) \text{ where } \hat{t}_i = f_\omega(x_i, z_i) \text{ (predict treatment)}$$

This determines $\hat{t}_i = f_\omega(x_i, z_i)$ for Stage Two.

Stage Two

$$\beta \in \arg\min_{\beta} \sum_i loss(y_i, \hat{y}_i) \text{ where } \hat{y}_i = g_\beta(x_i, \hat{t}_i) \text{ (predict outcome)}$$

$$\text{s.t.} \quad \beta \quad \text{obeys } \sum_i loss(r_i, \hat{r}_i) \geq \sum_i loss(r_i, 0) - \epsilon'$$

(The model cannot predict the remainder too much better than a zero model)

$$\text{where} \quad r_i = y_i - \hat{y}_i \text{ (remainder)}$$

$$\text{and} \quad \hat{r}_i = h_\alpha(x_i, z_i), \text{ where } \alpha \in \arg\min_{\alpha} \sum_i loss(r_i, \hat{r}_i) \text{ (predict remainder).}$$

4.1.2 Vectorized Version with General Loss

All quantities are vectorized in this version.

Stage One

$$\omega \in \arg\min_{\omega} \ loss(\vec{t}, \hat{t}) \text{ where } \hat{t} = f_\omega(\vec{X}, \vec{Z})$$

Stage Two

$$\begin{aligned} \beta \in \arg\min_{\beta} \ & loss(\vec{y}, \hat{y}) \text{ where } \hat{y} = g_\beta(\vec{X}, \hat{t}) \\ \text{s.t.} \quad \beta \quad & \text{obeys } loss(\vec{r}, \hat{r}) \geq loss(\vec{r}, 0) - \epsilon' \\ \text{where} \quad \vec{r} \quad & = \vec{y} - \hat{y} \\ \text{and} \quad \hat{r} \quad & = h_\alpha(\vec{X}, \vec{Z}), \text{ where } \alpha \in \arg\min_{\alpha} \ loss(\vec{r}, \hat{r}). \end{aligned}$$

4.1.3 Vectorized Version with Squared Loss

Note that we replaced the general loss function with the squared loss function in this version and analyze it in the remainder of this paper.

Stage One

$$\omega \in \arg\min_{\omega} (\vec{t} - \hat{t})^T (\vec{t} - \hat{t}) \text{ where } \hat{t} = f_\omega(\vec{X}, \vec{Z})$$

Stage Two

$$\begin{aligned} \beta \in \arg\min_{\beta} \ & (\vec{y} - \hat{y})^T (\vec{y} - \hat{y}) \text{ where } \hat{y} = g_\beta(\vec{X}, \hat{t}) \\ \text{s.t.} \quad \beta \quad & \text{obeys } (\vec{r} - \hat{r})^T (\vec{r} - \hat{r}) \geq (\vec{r} - 0)^T (\vec{r} - 0) - \epsilon' \\ \text{where} \quad \vec{r} \quad & = \vec{y} - \hat{y} \\ \text{and} \quad \hat{r} \quad & = h_\alpha(\vec{X}, \vec{Z}), \text{ where } \alpha \in \arg\min_{\alpha} (\vec{r} - \hat{r})^T (\vec{r} - \hat{r}). \end{aligned}$$

4.1.4 Simplified Version

Note that we solved α using the squared loss functions, $\alpha = (X_t^T X_t)^{-1} X_t^T r$ where $\vec{r} = \vec{y} - \hat{y}$, $\hat{y} = X_y \beta$ and eliminate it in this version.

Stage One

$$\omega \in \arg\min_{\omega} \ \omega^T X_t^T X_t \omega - 2\omega^T X_t^T \vec{t} + \vec{t}^T \vec{t} \tag{4.1}$$

Stage Two

$$\beta^T X_y^T X_y \beta - 2\beta^T X_y^T \vec{y} + \vec{y}^T \vec{y} \tag{4.2}$$

$$\text{s.t.} \quad \beta \quad \text{obeys } \beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H \vec{y} + \vec{y}^T H \vec{y} \leq \epsilon'. \tag{4.3}$$

where $\epsilon' = \gamma \tilde{r}^T \tilde{r}$ and $\tilde{r}$ is obtained by the traditional two stage method without constraints i.e.

$$\tilde{r} = \vec{y} - H_1 \vec{y} = \vec{y} - X_y(\vec{X}, H\vec{t})(X_y(\vec{X}, H\vec{t})^T X_y(\vec{X}, H\vec{t}))^{-1} X_y(\vec{X}, H\vec{t}))^T \vec{y}.$$

where $H = X_r(X_r^T X_r)^{-1} X_r^T$ and $H_1 = X_y(X_y^T X_y)^{-1} X_y^T$.

Also $X_t = X_t(\vec{X}, \vec{Z})$, $X_y = X_y(\vec{X}, \hat{t})$, and $X_r = X_r(\vec{X}, \vec{Z})$ are the predictor matrices of $\vec{t}$, $\vec{y}$, and $\vec{r}$ respectively. Also, $\hat{t} = X_t \hat{\omega}$.

The first stage is the same as that of the traditional two stage method when f_ω is a linear model with coefficients ω, and the loss is the squared loss. The second stage has a model for outcomes $\hat{y}$ depending on $\hat{t}$ and the covariates $\vec{X}$ as the traditional method does, however, the model is constrained. The constraint says that if we were to use the instrument to model the remainder (generating a model $\hat{r}$ to fit $\vec{r}$) then despite our efforts, we cannot predict the remainder much better than if we had used a model that was identically 0.

In this paper, we assume that ω, β, and α are coefficients for the general additive model. f_ω, g_β, and h_α are least squares models that estimate the treatment $\vec{t}$, the outcome $\vec{y}$, and the remainder $\vec{r}$ respectively.

4.2 One Stage Method

While this new two-stage formulation can help us to answer the questions stated in the introduction, it is possible that the constraints may not be obeyed in the second stage because of an incorrect model in the first stage. Often there are many models that predict

almost equally well on a finite dataset, and it is not clear exactly what the first stage model should be. It is possible that models that predict well in the first stage lead to residuals that can be predicted by the instrument in the second stage. In the two stage setting, there is no mechanism to change the first stage model after it is constructed in the first stage. The one stage formulation we will present next prevents this from happening. The formulation uses the notion of the "Rashomon set" that is, the set of models for with loss less than ϵ.

The first stage is replaced with a constraint that says any model $\hat{t}$ is feasible if it predicts $\vec{t}$ well, that is, it is in the Rashomon set. This is equivalent in the Bayesian setting to forcing a high posterior for $\hat{t}$.

4.2.1 Original Version

$$\min_{\beta,\omega} \sum_i loss(y_i, \hat{y}_i) \text{ where } \hat{y}_i = g_\beta(x_i, \hat{t}_i), \text{ and } \hat{t}_i = f_\omega(x_i, z_i) \text{ (predict outcome)}$$

$$\text{s.t.} \quad \omega \quad \text{obeys } \sum_i loss(t_i, \hat{t}_i) \leq \epsilon$$

(The model can predict the treatment well enough so that it is in the Rashomon set.)

$$\text{and} \quad \beta \quad \text{obeys } \sum_i loss(r_i, \hat{r}_i) \geq \sum_i loss(r_i, 0) - \epsilon'$$

(The model cannot predict the remainder too much better than a zero model)

$$\text{where} \quad r_i = y_i - \hat{y}_i \text{ (remainder)}$$

$$\text{and} \quad \hat{r}_i = h_\alpha(x_i, z_i), \text{ where } \alpha \in \arg\min_\alpha \sum_i loss(r_i, \hat{r}_i) \text{ (predict remainder).}$$

4.2.2 Vectorized Version with General Loss

Note that we simply conducted vectorization in this version.

$$\begin{aligned}
&\min_{\beta,\omega} \ loss(\vec{y},\hat{y}) \text{ where } \hat{y} = g_\beta(\vec{X},\hat{t}), \text{ and } \hat{t} = f_\omega(\vec{X},\vec{Z}) \\
\text{s.t.} \quad &\omega \quad \text{obeys } loss(\vec{t},\hat{t}) \leq \epsilon \\
\text{and} \quad &\beta \quad \text{obeys } loss(\vec{r},\hat{r}) \geq loss(\vec{r},0) - \epsilon' \\
\text{where} \quad &\vec{r} \quad = \vec{y} - \hat{y} \\
\text{and} \quad &\hat{r} \quad = h_\alpha(\vec{X},\vec{Z}), \text{ where } \alpha \in \arg\min_\alpha loss(\vec{r},\hat{r}).
\end{aligned}$$

4.2.3 Vectorized Version with Squared Loss

Note that we replaced the general loss function with the squared loss function in this version and analyze it in the remainder of this paper.

$$\begin{aligned}
&\min_{\beta,\omega} \ (\vec{y}-\hat{y})^T(\vec{y}-\hat{y}) \text{ where } \hat{y} = g_\beta(\vec{X},\hat{t}), \text{ and } \hat{t} = f_\omega(\vec{X},\vec{Z}) \\
\text{s.t.} \quad &\omega \quad \text{obeys } (\vec{t}-\hat{t})^T(\vec{t}-\hat{t}) \leq \epsilon \\
\text{and} \quad &\beta \quad \text{obeys } (\vec{r}-\hat{r})^T(\vec{r}-\hat{r}) \geq (\vec{r}-0)^T(\vec{r}-0) - \epsilon' \\
\text{where} \quad &\vec{r} \quad = \vec{y} - \hat{y} \\
\text{and} \quad &\hat{r} \quad = h_\alpha(\vec{X},\vec{Z}), \text{ where } \alpha \in \arg\min_\alpha (\vec{r}-\hat{r})^T(\vec{r}-\hat{r}).
\end{aligned}$$

4.2.4 Simplified Version

Note that we solved α using the squared loss functions, $\alpha = (X_t^T X_t)^{-1} X_t^T \vec{r}$ where $\vec{r} = \vec{y} - \hat{y}$, $\hat{y} = X_y\beta$ and eliminate it in this version.

$$\min_{\beta,\omega} \ \beta^T X_y^T X_y \beta - 2\beta^T X_y^T \vec{y} + \vec{y}^T\vec{y} \tag{4.4}$$

$$\text{s.t.} \quad \omega \quad \text{obeys } \omega^T X_t^T X_t \omega - 2\omega^T X_t^T \vec{t} + \vec{t}^T\vec{t} \leq \epsilon \tag{4.5}$$

$$\text{and} \quad \beta \quad \text{obeys } \beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon' \tag{4.6}$$

Here, $\epsilon = \gamma\vec{t}^T\vec{t}$, $\epsilon' = \gamma\tilde{r}^T\tilde{r}$, and $H = X_r(X_r^T X_r)^{-1}X_r^T$. Also $\tilde{r}$ is obtained by the

traditional two stage method without constraints i.e.

$$\tilde{r} = \vec{y} - X_y(X_y^T X_y)^{-1} X_y^T \vec{y}.$$

Also, $X_t = X_t(\vec{X}, \vec{Z})$ and $X_y = X_y(\vec{X}, X_t\omega)$ are the predictor matrices of $\vec{t}$ and $\vec{y}$ respectively.

Chapter 5

Summary

The tables below provide a summary of potential optimization results using our two stage method under different model constructions in the uni-dimensional and multi-dimensional cases of the instrument respectively, which illustrates the limitations of our two stage method under specific settings.

Table 5.1: Potential Optimization Results under Different Model Constructions in the Uni-dimensional Cases of Instrument

Prediction Model in Stage One	**Prediction Model in Stage Two**	**Potential Optimization Results in $1-D$ Cases**
LM	LM	The constraint is always satisfied (or never active). (Theorem 2)
LM	GAM	The constraint is always satisfied (or never active). (Theorem 3)
GAM	LM	EITHER The constraints is not active OR There is no feasible solution (Theorem 1)
GAM	GAM	EITHER The constraints is not active OR There is no feasible solution OR There exists a feasible solution

Table 5.2: Potential Optimization Results under Different Model Constructions in the Multi-dimensional Cases of Instrument

Prediction Model in Stage One	**Prediction Model in Stage Two**	**Potential Optimization Results in 2^+-D Cases**
LM	LM	EITHER The constraints is not active OR There is no feasible solution (Theorem 1)
LM	GAM	EITHER The constraints is not active OR There is no feasible solution OR There exists a feasible solution
GAM	LM	EITHER The constraints is not active OR There is no feasible solution (Theorem 1)

GAM	GAM	EITHER The constraints is not active OR There is no feasible solution OR There exists a feasible solution

Note: 1. LM stands for linear model and GAM stands for general additive model. 2. Only the prediction models in Stage One and Stage Two will influence the potential optimization results, but the data generation models (or the true models) will not influence them. 3. Theorem 2 is a special case of Theorem 1.

The table below provides an intuitive explanation of each potential optimization results using our two stage method.

Table 5.3: Potential Optimization Results and Intuitions

Optimization Result	**Intuition**
The constraint is never active.	The validity check is not useful and we cannot check whether the instrument is good or not.
The constraint can be active, but there is no feasible solution.	The validity check is not satisfied, which means the instrument is not good. And we cannot fix it.
The constraint can be active, and there can be a feasible solution.	The validity check is not satisfied, which means the instrument is not good. But we can to some extent improve it.

Chapter 6

Lemmas and Theorems

Note that all the theorems in this chapter rely on the assumption that the flexible model classes for the treatment t and for the remainder r have the same level of complexity, that is, predictor matrices for the treatment t and the remainder r are equivalent i.e. $Xt = Xr$. Therefore, the hat matrix of the remainder r, $H = X_r(X_r^T X_r)^{-1} X_r^T$ can be written as $H = X_r(X_r^T X_r)^{-1} X_r^T = X_t(X_t^T X_t)^{-1} X_t^T$, which is the hat matrix for the treatment t.

In the remainder of this section, we use the definition $H = X_t(X_t^T X_t)^{-1} X_t^T$ to represent both hat matrices for the remainder r and the treatment t.

6.1 Notations

The notations used in the remainder of this chapter are illustrated as below.

The covariates $\vec{X}$ is a $n \times p$ matrix whose column space is $p-$dimensional i.e. $\vec{X} = (x_1, \cdots, x_p)$.

In the multi-dimensional cases, the instrument $\vec{Z}$ is a $n \times q$ matrix whose column space is $q-$dimensional i.e. $\vec{Z} = (z_1, \cdots, z_q)$.

In the uni-dimensional cases, the instrument $\vec{Z}$ is a $n \times 1$ matrix whose column space is $1-$dimensional i.e. $\vec{Z} = (z_1)$.

Stage One

The predictor matrix of $\vec{t}$ in Stage One is $X_t = X_t(\vec{X}, \vec{Z})$.

The hat matrix of $\vec{t}$ in Stage One is $H = X_t(X_t^T X_t)^{-1} X_t^T$.

The predicted value of $\vec{t}$ in Stage One is $\hat{t} = X_t\hat{\omega}$ and $\hat{\omega} = (X_t^T X_t)^{-1} X_t^T \vec{t}$. Thus, $\hat{t} = X_t\hat{\omega} = X_t(X_t^T X_t)^{-1} X_t^T \vec{t} = H\vec{t}$

Stage Two

The predictor matrix of $\vec{y}$ in Stage Two is $X_y = X_y(\vec{X}, \hat{t})$, where $\hat{t} = X_t\hat{\omega}$.

The hat matrix of $\vec{y}$ in Stage Two is $H_1 = X_y(X_y^T X_y)^{-1} X_y^T$.

6.2 $2^+ - D$ Instrument

Linear Algebra Theorem. *A triangular matrix is invertible, if and only if all of its diagonal entries are nonzero.*

Lemma 1. *If the models in Stage One and Stage Two are both linear models, the optimal / minimum solution $\hat{\beta}_{min}$ of the objective function (2) ($\beta^T X_y^T X_y \beta - 2\beta^T X_y^T \vec{y} + \vec{y}^T \vec{y}$) equals to the optimal / minimum solution $\hat{\beta}'_{min}$ of the objective function on the left side of the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y}$) i.e. $\hat{\beta}_{min} = \hat{\beta}'_{min}$. This statement is true regardless of the dimension of $\vec{X}$ and the dimension of $\vec{Z}$.*

Proof. Recall the notation for the model in Stage One is: $\hat{t} = f_\omega(\vec{X}, \vec{Z}) = f_\omega(x_1, \cdots, x_p, z_1, \cdots, z_q)$. To notate that it is a linear model, we use the following notation: $\hat{t} = \omega_1 x_1 + \cdots + \omega_p x_p + \omega_{p+1} z_1 + \cdots + \omega_{p+q} z_q$.

Recall notation for the model in Stage Two is: $\hat{y} = g_\beta(\vec{X}, \hat{t}) = g_\beta(x_1, \cdots, x_p, \hat{t})$. To notate that it is a linear model, we use the following notation: $\hat{y} = \beta_1 x_1 + \cdots + \beta_p x_p + \beta_{p+1}\hat{t}$.

Define the predictor matrix of t in Stage One:

$$X_t = (x_1, \cdots, x_p, z_1, \cdots, z_q)_{n\times(p+q)} \in \mathbb{R}^{n\times(p+q)}.$$

Define the predictor matrix of y in Stage Two:

$$
\begin{aligned}
X_y =&(x_1, \cdots, x_p, \hat{t})_{n\times(p+1)} \\
=&(x_1, \cdots, x_p, z_1, \cdots, z_q)_{n\times(p+q)} \begin{pmatrix} 1 & \cdots & 0 & \omega_1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 1 & \omega_p \\ 0 & \cdots & 0 & \omega_{p+1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \omega_{p+q} \end{pmatrix}_{(p+q)\times(p+1)} \\
=:&X_t B \in \mathbb{R}^{n\times(p+1)}.
\end{aligned}
$$

In Stage Two of our two stage method, the optimal / minimum solution of the objective function (2) ($\beta^T X_y^T X_y \beta - 2\beta^T X_y^T \vec{y} + \vec{y}^T \vec{y}$) is $\hat{\beta}_{min} = [X_y^T X_y]^{-1} X_y^T \vec{y}$, and the optimal / minimum solution of the objective function of the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y}$) is $\hat{\beta}'_{min} = [X_y^T H X_y]^{-1} X_y^T H\vec{y}$. Then,

$$
\begin{aligned}
\hat{\beta}'_{min} =&[X_y^T H X_y]^{-1} X_y^T H\vec{y} \\
=&[B^T X_t^T H X_t B]^{-1} B^T X_t^T H\vec{y}, \text{ where } X_y = X_t B \\
=&[B^T X_t^T X_t [X_t^T X_t]^{-1} X_t^T X_t B]^{-1} B^T X_t^T X_t [X_t^T X_t]^{-1} X_t^T \vec{y}, \text{ where } X_y = X_t B \\
=&[B^T X_t^T X_t B]^{-1} B^T X_t^T \vec{y} \\
=&\hat{\beta}_{min}.
\end{aligned}
$$

Therefore, the two optimal solutions are equivalent, i.e., $\hat{\beta}_{min} = \hat{\beta}'_{min}$.

Lemma 2. *If the model in Stage One is a general additive model, and the model in Stage Two is a linear model, Lemma 1 still holds. This statement is true regardless of the dimension of $\vec{X}$ and the dimension of $\vec{Z}$.*

Proof. Recall the notation for the model in Stage One is: $\hat{t} = f_\omega(\vec{X}, \vec{Z}) = f_\omega(x_1, \cdots, x_p, z_1, \cdots, z_q)$. To notate that it is a general additive model, we use the following notation: $\hat{t} = \omega_1 x_1 + \cdots + \omega_p x_p + \omega_{p+1} z_1 + \cdots + \omega_{p+q} z_q + \omega_{p+q+1} \text{feature}_1(\vec{X}, \vec{Z}) + \cdots + \omega_{p+q+k} \text{feature}_k(\vec{X}, \vec{Z})$, where $\text{feature}_j(\vec{X}, \vec{Z})$, $j = 1, \cdots, k$ is a non-linear function of $(\vec{X}, \vec{Z})$.

Recall notation for the model in Stage Two is: $\hat{y} = g_\beta(\vec{X}, \hat{t}) = g_\beta(x_1, \cdots, x_p, \hat{t})$. To notate that it is a linear model, we use the following notation: $\hat{y} = \beta_1 x_1 + \cdots + \beta_p x_p + \beta_{p+1}\hat{t}$.

Define the predictor matrix of t in Stage One:

$$X_t = (x_1, \cdots, x_p, z_1, \cdots, z_q, \text{feature}_1(\vec{X}, \vec{Z}), \cdots, \text{feature}_k(\vec{X}, \vec{Z}))_{n\times(p+q+k)} \in \mathbb{R}^{n\times(p+q+k)}$$

Define the predictor matrix of y in Stage Two:

$$\begin{aligned}
X_y =& (x_1, \cdots, x_p, \hat{t})_{n\times(p+1)} \\
=& (x_1, \cdots, x_p, z_1, \cdots, z_q, \text{feature}_1(\vec{X}, \vec{Z}), \cdots, \text{feature}_k(\vec{X}, \vec{Z}))_{n\times(p+q+k)} \\
& \times \begin{pmatrix} 1 & \cdots & 0 & \omega_1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 1 & \omega_p \\ 0 & \cdots & 0 & \omega_{p+1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \omega_{p+q} \\ 0 & \cdots & 0 & \omega_{p+q+1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \omega_{p+q+k} \end{pmatrix}_{(p+q+k)\times(p+1)} \\
=:& X_t B' \in \mathbb{R}^{n\times(p+1)}
\end{aligned}$$

Since the matrix B' has the same form as the matrix B in Lemma 1, it has already been proved that two optimal / minimum solutions are equivalent i.e. $\hat{\beta}_{min} = \hat{\beta}'_{min}$.

Lemma 3. *If the models in Stage One and Stage two are general additive models that have the following forms.*

$$\begin{aligned}
\hat{t} =& \omega_1 \text{feature}_1(\vec{X}) + \cdots + \omega_{k_1}\text{feature}_{k_1}(\vec{X}) + \\
& \omega_{k_1+1}\text{feature}_{k_1+1}(\vec{X}, \vec{Z}) + \cdots + \omega_{k_1+k_2}\text{feature}_{k_1+k_2}(\vec{X}, \vec{Z})
\end{aligned}$$

where $\text{feature}_j(\vec{X})$, $j = 1, \cdots, k_1$ *is a linear or non-linear function of* $\vec{X}$ *and* $\text{feature}_j(\vec{X}, \vec{Z})$, $j = k_1 + 1, \cdots, k_1 + k_2$ *is a linear or non-linear function of* $(\vec{X}, \vec{Z})$ *or only* $\vec{Z}$.

$$\hat{y} = \beta_1 \text{feature}_1(\vec{X}) + \cdots + \beta_{k'_1}\text{feature}_{k'_1}(\vec{X}) + \beta_{k'_1+1}\hat{t}$$

where $k_1' \leq k_1$ *and* $\{\text{feature}_j(\vec{X})\}_{j=1}^{k_1'}$ *is a subset of* $\{\text{feature}_j(\vec{X})\}_{j=1}^{k_1}$.

In other word, the following conditions are satisfied:

1. the input features of the covariates x in Stage Two is a subset of that in Stage One;

2. the model in Stage Two only contains the linear term of the predicted values of the treatment $\hat{t}$.

then Lemma 1 still holds. This statement is true regardless of the dimension of $\vec{X}$ *and the dimension of* $\vec{Z}$. *Note that Lemma 3. is a more general version of both Lemma 1. and Lemma 2.*

Proof. Recall the notation of the models in Stage One and Stage Two in **Lemma 3.** and define the predictor matrices as follows:

Define the predictor matrix of t in Stage One:

$$X_t = (\text{feature}_1(\vec{X}), \cdots, \text{feature}_{k_1}(\vec{X}), \text{feature}_{k_1+1}(\vec{X}, \vec{Z}), \cdots, \text{feature}_{k_1+k_2}(\vec{X}, \vec{Z}))_{n\times(k_1+k_2)} \in \mathbb{R}^{n\times(k_1+k_2)}$$

Define the predictor matrix of y in Stage Two:

$$\begin{aligned} X_y =& (\text{feature}_1(\vec{X}), \cdots, \text{feature}_{k_1'}(\vec{X}), \hat{t})_{n\times(p+1)} \\ =& (\text{feature}_1(\vec{X}), \cdots, \text{feature}_{k_1}(\vec{X}), \text{feature}_{k_1+1}(\vec{X}, \vec{Z}), \cdots, \text{feature}_{k_1+k_2}(\vec{X}, \vec{Z}))_{n\times(k_1+k_2)} \\ & \times \begin{pmatrix} 1 & \cdots & 0 & \omega_1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 1 & \omega_{k_1'} \\ 0 & \cdots & 0 & \omega_{k_1'+1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \omega_{k_1} \\ 0 & \cdots & 0 & \omega_{k_1+1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \omega_{k_1+k_2} \end{pmatrix}_{(k_1+k_2)\times(k_1'+1)} \\ =:& X_t B'' \in \mathbb{R}^{n\times(k_1'+1)} \end{aligned}$$

Since the matrix B'' has the same form as the matrix B in Lemma 1, it has already been proved that two optimal / minimum solutions are equivalent i.e. $\hat{\beta}_{min} = \hat{\beta}'_{min}$.

Theorem 1. *If the model in Stage Two is a linear model, and if the model in Stage One is a linear model or a general additive model, either the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon'$) is not active or there is no feasible solution. This statement is true regardless of the dimension of $\vec{X}$ and the dimension of $\vec{Z}$.*

Proof. To notate the objective function (2) ($\beta^T X_y^T X_y \beta - 2\beta^T X_y^T \vec{y} + \vec{y}^T \vec{y}$), we use the following notation: $C_\beta(x_1, \cdots, x_p, \hat{t})$.

To notate the objective function on the left side of the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y}$), we use the following notation: $C_{\beta'}(x_1, \cdots, x_p, \hat{t})$.

Recall **Lemma 1.** and **Lemma 2.**, the optimal / minimum solution of the objective function (2) ($C_\beta(x_1, \cdots, x_p, \hat{t})$) is $\hat{\beta}_{min}$, and the optimal / minimum solution of the objective function on the left side of the constraint (3) ($C_{\beta'}(x_1, \cdots, x_p, \hat{t})$) is $\hat{\beta}'_{min}$. Two optimal / minimum solutions are equivalent i.e. $\hat{\beta}_{min} = \hat{\beta}'_{min}$.

If the constraint (3) is active, then $\hat{\beta}_{min}$ does not satisfy it i.e. $C_{\hat{\beta}_{min}}(x_1, \cdots, x_p, \hat{t}) > \epsilon'$.

$min\ C_{\beta'}(x_1, \cdots, x_p, \hat{t})$

$= C_{\hat{\beta}'_{min}}(x_1, \cdots, x_p, \hat{t})$, where $\hat{\beta}'_{min}$ is the optimal / minimum solution of $C_{\beta'}(x_1, \cdots, x_p, \hat{t})$

$= C_{\hat{\beta}_{min}}(x_1, \cdots, x_p, \hat{t}) > \epsilon'$, where $\hat{\beta}'_{min} = \hat{\beta}_{min}$

Therefore, $C_{\beta'}(x_1, \cdots, x_p, \hat{t}) > \epsilon'$, the constraint (3) is never satisfied. In conclusion, once the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon'$) is active, there does not exist a solution that can satisfy it.

More General Version of Theorem 1. *If the models in Stage One and Stage two are general additive models that have the following forms.*

$$\begin{aligned}\hat{t} =& \omega_1 \text{feature}_1(\vec{X}) + \cdots + \omega_{k_1} \text{feature}_{k_1}(\vec{X}) + \\ & \omega_{k_1+1} \text{feature}_{k_1+1}(\vec{X}, \vec{Z}) + \cdots + \omega_{k_1+k_2} \text{feature}_{k_1+k_2}(\vec{X}, \vec{Z})\end{aligned}$$

where $\text{feature}_j(\vec{X})$, $j = 1, \cdots, k_1$ *is a linear or non-linear function of* $\vec{X}$ *and* $\text{feature}_j(\vec{X}, \vec{Z})$,

$j = k_1 + 1, \cdots, k_1 + k_2$ is a linear or non-linear function of $(\vec{X}, \vec{Z})$ or only $\vec{Z}$.

$$\hat{y} = \beta_1 \text{feature}_1(\vec{X}) + \cdots + \beta_{k_1'} \text{feature}_{k_1'}(\vec{X}) + \beta_{k_1'+1}\hat{t}$$

where $k_1' \leq k_1$ and $\{\text{feature}_j(\vec{X})\}_{j=1}^{k_1'}$ is a subset of $\{\text{feature}_j(\vec{X})\}_{j=1}^{k_1}$.

In other word, the following conditions are satisfied:

1. the input features of the covariates x in Stage Two is a subset of that in Stage One;

2. the model in Stage Two only contains the linear term of the predicted values of the treatment $\hat{t}$.

then either the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon'$) is not active or there is no feasible solution. This statement is true regardless of the dimension of $\vec{X}$ and the dimension of $\vec{Z}$.

Proof. The proof of the **More General Version of Theorem 1.** using **Lemma 3.** is the same as the proof of **Theorem 1.** using **Lemma 1.** and **Lemma 2.**.

6.3 $1 - D$ Instrument

Theorem 2. *In the uni-dimensional case of the instrument $\vec{Z}$, if the models in Stage One and Stage Two are both linear models, the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon'$) is always satisfied (or never active). This statement is true regardless of the dimension of $\vec{X}$.*

Proof. Recall the notation for the model in Stage One is: $\hat{t} = f_\omega(\vec{X}, \vec{Z}) = f_\omega(x_1, \cdots, x_p, z_1)$. To notate that it is a linear model, we use the following notation: $\hat{t} = \omega_1 x_1 + \cdots + \omega_p x_p + \omega_{p+1} z_1$.

Recall notation for the model in Stage Two is: $\hat{y} = g_\beta(\vec{X}, \hat{t}) = g_\beta(x_1, \cdots, x_p, \hat{t})$. To notate that it is a linear model, we use the following notation: $\hat{y} = \beta_1 x_1 + \cdots + \beta_p x_p + \beta_{p+1}\hat{t}$.

Define the predictor matrix of t in Stage One:

$$X_t = (x_1, x_2, \cdots, x_p, z_1)_{n \times (p+1)} \in \mathbb{R}^{n \times (p+1)}$$

Define the predictor matrix of y in Stage Two:

$$
\begin{aligned}
X_y =& (x_1, x_2 \cdots, x_p, \hat{t})_{n\times(p+1)} \\
=& (x_1, x_2, \cdots, x_p, z_1)_{n\times(p+1)} \begin{pmatrix} 1 & 0 & \cdots & 0 & \omega_1 \\ 0 & 1 & \cdots & 0 & \omega_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \omega_p \\ 0 & 0 & \cdots & 0 & \omega_{p+1} \end{pmatrix}_{(p+1)\times(p+1)} \\
=:& X_t A \in \mathbb{R}^{n\times(p+1)}
\end{aligned}
$$

where diagonal entries of A are $a_{ii} = 1$, $1 \leq i \leq p$ and $a_{(p+1)(p+1)} = \omega_{p+1}$. Therefore, the upper trapezoidal matrix A has non-zero diagonal entries.

As shown in the following, the hat matrix H_1 equals to the hat matrix H.

$$
\begin{aligned}
H_1 &= X_y[X_y^T X_y]^{-1} X_y^T, \text{ where } X_y = X_t A \\
&= X_t A[A^T X_t^T X_t A]^{-1} A^T X_t^T, \text{ where } A \in \mathbb{R}^{(p+1)\times(p+1)} \\
&= X_t A A^{-1}[X_t^T X_t]^{-1} A^{-T} A^T X_t^T \\
&= X_t[X_t^T X_t]^{-1} X_t^T \\
&= H
\end{aligned}
$$

In Stage Two of our two stage method, the optimal / minimum solution of the objective function (2) ($\beta^T X_y^T X_y \beta - 2\beta^T X_y^T \vec{y} + \vec{y}^T \vec{y}$) is $\hat{\beta}_{min} = [X_y^T X_y]^{-1} X_y^T \vec{y}$ and $\hat{y} = X_y \hat{\beta}_{min} = X_y[X_y^T X_y]^{-1} X_y^T \vec{y} = H_1 \vec{y}$.

The objective function on the left side of the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H \vec{y} +$

$\vec{y}^T H\vec{y}$) is:

$$
\begin{aligned}
&||H(\vec{y}-\hat{y})||^2 \\
=&||H\vec{y}-H\hat{y}||^2 \text{ where } \hat{y}=H_1 y \\
=&||H\vec{y}-HH_1\vec{y}||^2 \text{ where } H_1=H \\
=&||H\vec{y}-H^2\vec{y}||^2 \text{ where } H^2=H \\
=&||H\vec{y}-H\vec{y}||^2 \\
=&0\leq\epsilon' \text{ where } \epsilon'>0
\end{aligned}
$$

Therefore, the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon'$) is always satisfied (or never active).

More General Version of Theorem 2. *In the uni-dimensional case of the instrument $\vec{Z}$, if the models in Stage One and Stage Two are general additive models that have the following forms.*

$$\hat{t} = \omega_1 \text{feature}_1(\vec{X}) + \cdots + \omega_k \text{feature}_k(\vec{X}) + \omega_{k+1}\text{feature}_{k+1}(\vec{Z})$$

where $\text{feature}_j(\vec{X})$, $j = 1, \cdots, k$ *is a linear or non-linear function of* $\vec{X}$ *and* $\text{feature}_{k+1}(\vec{Z})$ *is a linear or non-linear function of* $\vec{Z}$.

$$\hat{y} = \beta_1 \text{feature}_1(\vec{X}) + \cdots + \beta_k \text{feature}_k(\vec{X}) + \beta_{k+1}\hat{t}$$

where $\text{feature}_j(\vec{X})$, $j = 1, \cdots, k$ *is a linear or non-linear function of* $\vec{X}$.

In other word, the following conditions are satisfied:

1. the models in Stage One and Stage two share the same input features of the covariates $\vec{X}$;

2. the model in Stage One contains only one input feature of the instrument $\vec{Z}$, which can be either linear or non-linear;

3. the model in Stage One contains no interaction term of the covariates $\vec{X}$ and the instrument $\vec{Z}$;

4. the model in Stage Two only contains the linear term of the predicted values of the treatment $\hat{t}$.

then the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon'$) is always satisfied (or never active). This statement is true regardless of the dimension of $\vec{X}$.

Proof. Recall the notation of the models in Stage One and Stage Two in the **More General Version of Theorem 2.** and define the predictor matrices as follows:

Define the predictor matrix of t in Stage One:

$$X_t = (\text{feature}_1(\vec{X}), \cdots, \text{feature}_k(\vec{X}), \text{feature}_{k+1}(\vec{Z}))_{n\times(k+1)} \in \mathbb{R}^{n\times(k+1)}$$

Define the predictor matrix of y in Stage Two:

$$\begin{aligned} X_y =& (\text{feature}_1(\vec{X}), \cdots, \text{feature}_k(\vec{X}), \hat{t})_{n\times(k+1)} \\ =& (\text{feature}_1(\vec{X}), \cdots, \text{feature}_k(\vec{X}), \text{feature}_{k+1}(\vec{Z}))_{n\times(k+1)} \begin{pmatrix} 1 & \cdots & 0 & \omega_1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 1 & \omega_k \\ 0 & \cdots & 0 & \omega_{k+1} \end{pmatrix}_{(k+1)\times(k+1)} \\ =:& X_t A \in \mathbb{R}^{n\times(k+1)} \end{aligned}$$

where diagonal entries of A are $a_{ii} = 1$, $1 \leq i \leq k$ and $a_{(k+1)(k+1)} = \omega_{k+1}$. Therefore, the upper trapezoidal matrix A has non-zero diagonal entries.

As shown in the following, the hat matrix H_1 equals to the hat matrix H.

$$\begin{aligned} H_1 &= X_y[X_y^T X_y]^{-1} X_y^T, \text{ where } X_y = X_t A \\ &= X_t A[A^T X_t^T X_t A]^{-1} A^T X_t^T, \text{ where } A \in \mathbb{R}^{(k+1)\times(k+1)} \\ &= X_t A A^{-1}[X_t^T X_t]^{-1} A^{-T} A^T X_t^T \\ &= X_t[X_t^T X_t]^{-1} X_t^T \\ &= H \end{aligned}$$

In Stage Two of our two stage method, the optimal / minimum solution of the objective function (2) ($\beta^T X_y^T X_y \beta - 2\beta^T X_y^T \vec{y} + \vec{y}^T \vec{y}$) is $\hat{\beta}_{min} = [X_y^T X_y]^{-1} X_y^T \vec{y}$ and $\hat{y} = X_y \hat{\beta}_{min} = X_y[X_y^T X_y]^{-1} X_y^T \vec{y} = H_1 \vec{y}$.

The objective function on the left side of the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y}$) is:

$$
\begin{aligned}
&||H(\vec{y} - \hat{y})||^2 \\
=&||H\vec{y} - H\hat{y}||^2 \text{ where } \hat{y} = H_1 y \\
=&||H\vec{y} - HH_1 y||^2 \text{ where } H_1 = H \\
=&||H\vec{y} - H^2 y||^2 \text{ where } H^2 = H \\
=&||H\vec{y} - H\vec{y}||^2 \\
=&0 \leq \epsilon' \text{ where } \epsilon' > 0
\end{aligned}
$$

Therefore, the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon'$) is always satisfied (or never active).

Lemma 4. *If $A \in \mathbb{R}^{n \times m}$, $n > m$ is an upper trapezoidal matrix with non-zero diagonal entries, the matrix $A[A^T A]^{-1} A^T$ is a block matrix of the form $A[A^T A]^{-1} A^T = \begin{pmatrix} I_{m \times m} & 0_{m \times (n-m)} \\ 0_{(n-m) \times m} & 0_{(n-m) \times (n-m)} \end{pmatrix}$, $I_{m \times m}$ is an m-dimensional identity matrix.*

Proof. Denote $A = \begin{pmatrix} C_{m \times m} \\ 0_{(n-m) \times m} \end{pmatrix}$, where $C \in \mathbb{R}^{m \times m}$ is an upper triangular matrix

with non-zero diagonal entries.

$$
\begin{aligned}
& A[A^TA]^{-1}A^T \\
&= \begin{pmatrix} C_{m\times m} \\ 0_{(n-m)\times m} \end{pmatrix} [\begin{pmatrix} C^T_{m\times m} & 0_{(n-m)\times m} \end{pmatrix} \begin{pmatrix} C_{m\times m} \\ 0_{(n-m)\times m} \end{pmatrix}]^{-1} \begin{pmatrix} C^T_{m\times m} & 0_{(n-m)\times m} \end{pmatrix} \\
&= \begin{pmatrix} C_{m\times m} \\ 0_{(n-m)\times m} \end{pmatrix} [C^T_{m\times m}C_{m\times m}]^{-1} \begin{pmatrix} C^T_{m\times m} & 0_{(n-m)\times m} \end{pmatrix} \\
&= \begin{pmatrix} C_{m\times m} \\ 0_{(n-m)\times m} \end{pmatrix} C^{-1}_{m\times m}C^{-T}_{m\times m} \begin{pmatrix} C^T_{m\times m} & 0_{(n-m)\times m} \end{pmatrix} \\
&= \begin{pmatrix} I_{m\times m} \\ 0_{(n-m)\times m} \end{pmatrix} \begin{pmatrix} I_{m\times m} & 0_{(n-m)\times m} \end{pmatrix} \\
&= \begin{pmatrix} I_{m\times m} & 0_{m\times(n-m)} \\ 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)} \end{pmatrix}
\end{aligned}
$$

Lemma 5. *If $B \in \mathbb{R}^{n\times n}$ is an upper triangular matrix with non-zero diagonal entries, the matrix $B[B^TB]^{-1}B^T = I_{n\times n}$ is an n-dimensional identity matrix.*

Proof.

$$B[B^TB]^{-1}B^T = BB^{-1}B^{-T}B^T = I_{n\times n}I_{n\times n} = I_{n\times n}$$

Theorem 3. *In the uni-dimensional case for the instrument $\vec{Z}$, if the model in Stage One is a linear model, and the model in Stage Two is a general additive model, the constraint (3) ($\beta^TX_y^THX_y\beta - 2\beta^TX_y^TH\vec{y} + \vec{y}^TH\vec{y} \leq \epsilon'$) is also never active. This statement is true regardless of the dimension of $\vec{X}$.*

Proof. Recall that **Theorem 2.** holds because of $H = H_1$. In **Theorem 3.**, although $H \neq H_1$, $H - HH_1 = 0$.

In Stage Two of our method, the optimal / minimum solution of the objective function (2) ($\beta^TX_y^TX_y\beta - 2\beta^TX_y^T\vec{y} + \vec{y}^T\vec{y}$) is $\hat{\beta}_{min} = [X_y^TX_y]^{-1}X_y^T\vec{y}$ and $\hat{y} = X_y\hat{\beta}_{min} = X_y[X_y^TX_y]^{-1}X_y^T\vec{y} = H_1\vec{y}$.

The objective function on the left side of the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y}$) is:

$$\begin{aligned}
&||H(\vec{y} - \hat{y})||^2 \\
=&||H\vec{y} - H\hat{y})||^2 \\
=&||H\vec{y} - HH_1 y)||^2 \text{ where } \hat{y} = H_1 y \\
=&||(H - HH_1)y)||^2 \text{ where } H - HH_1 = 0 \\
=&0 \leq \epsilon' \text{ where } \epsilon' > 0
\end{aligned}$$

Therefore, the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H\vec{y} + \vec{y}^T H\vec{y} \leq \epsilon'$) is always satisfied (or never active).

Next prove $H - HH_1 = 0$.

Recall the notation for the model in Stage One is: $\hat{t} = f_\omega(\vec{X}, \vec{Z}) = f_\omega(x_1, \cdots, x_p, z_1)$. To notate that it is a linear model, we use the following notation: $\hat{t} = \omega_1 x_1 + \cdots + \omega_p x_p + \omega_{p+1} z_1$.

Recall the notation for the model in Stage Two is: $\hat{y} = g_\beta(\vec{X}, \hat{t}) = g_\beta(x_1, \cdots, x_p, \hat{t})$. To notate that it is a general additive model, we use the following notation: $\hat{y} = \beta_1 x_1 + \cdots + \beta_p x_p + \beta_{p+1}\hat{t} + \beta_{p+2}\text{feature}_1(\vec{X}, \hat{t}) + \cdots + \beta_{p+k+1}\text{feature}_k(\vec{X}, \hat{t})$, where $\text{feature}_j(\vec{X}, \hat{t})$, $j = 1, \cdots, k$ is a non-linear function of $(\vec{X}, \hat{t})$.

Define the predictor matrix of t in Stage One:

$$X_t = (x_1, \cdots, x_p, z_1) \in \mathbb{R}^{n \times (p+1)}.$$

Define the predictor matrix of y in Stage Two:

$$X_y = (x_1, \cdots, x_p, \hat{t}, \text{feature}_1(\vec{X}, \hat{t}), \cdots, \text{feature}_k(\vec{X}, \hat{t})) \in \mathbb{R}^{n \times m}$$

where $m = p + k + 1$.

Use the Gram-Schmidt algorithm to construct an orthogonal set of unit vectors.

Step 1: $u_1 = x_1$, $e_1 = \frac{u_1}{|u_1|}$

Step 2: $u_2 = x_2 - \frac{x_2 \cdot u_1}{|u_1|^2} u_1$, $e_2 = \frac{u_2}{|u_2|}$

$\vdots$

Step p+1: $u_{p+1} = \hat{t} - \frac{\hat{t}\cdot u_1}{|u_1|^2}u_1 - \cdots - \frac{\hat{t}\cdot u_p}{|u_p|^2}u_p$, $e_{p+1} = \frac{u_{p+1}}{|u_{p+1}|}$

Step p+2: $u_{p+2} = \text{feature}_1(\vec{X}, \hat{t}) - \frac{\text{feature}_1(\vec{X},\hat{t})\cdot u_1}{|u_1|^2}u_1 - \cdots - \frac{\text{feature}_1(\vec{X},\hat{t})\cdot u_{p+1}}{|u_{p+1}|^2}u_{p+1}$, $e_{p+2} = \frac{u_{p+2}}{|u_{p+2}|}$

$\vdots$

Step m: $u_m = \text{feature}_k(\vec{X}, \hat{t}) - \frac{\text{feature}_k(\vec{X},\hat{t})\cdot u_1}{|u_1|^2}u_1 - \cdots - \frac{\text{feature}_k(\vec{X},\hat{t})\cdot u_{m-1}}{|u_{m-1}|^2}u_{m-1}$, $e_m = \frac{u_m}{|u_m|}$

Therefore the predictor matrix and t and y can be written as follows,

$$
\begin{aligned}
X_t =& (x_1, \cdots, x_p, z_1) \\
=& (u_1, u_2, \cdots, u_m) \begin{pmatrix} 1 & a_{12} & \cdots & a_{1(p+1)} \\ 0 & 1 & \cdots & a_{2(p+1)} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\omega_{p+1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}_{m\times(p+1)} \\
=:& UA_u = (e_1, e_2, \cdots, e_m) \begin{pmatrix} |u_1| & 0 & \cdots & 0 \\ 0 & |u_2| & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & |u_m| \end{pmatrix}_{m\times m} \begin{pmatrix} 1 & a_{12} & \cdots & a_{1(p+1)} \\ 0 & 1 & \cdots & a_{2(p+1)} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\omega_{p+1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}_{m\times(p+1)} \\
=:& ED_uA_u \\
=:& EA
\end{aligned}
$$

where diagonal entries of D_u are $d_{ii} = |u_i| > 0$, and diagonal entries of A_u are $a_{ii} = 1$, $1 \le i \le p$ and $a_{(p+1)(p+1)} = 1/\omega_{p+1}$. Therefore, the upper trapezoidal matrix $A = D_uA_u$

has non-zero diagonal entries.

$$
\begin{aligned}
X_y =&(x_1, \cdots, x_p, \hat{t}, \text{feature}_1(\vec{X}, \hat{t}), \cdots, \text{feature}_k(\vec{X}, \hat{t})) \\
=&(u_1, u_2, \cdots, u_m) \begin{pmatrix} 1 & b_{12} & \cdots & b_{1m} \\ 0 & 1 & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{m \times m} \\
=:&UB_u = (e_1, e_2, \cdots, e_m) \begin{pmatrix} |u_1| & 0 & \cdots & 0 \\ 0 & |u_2| & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & |u_m| \end{pmatrix}_{m \times m} \begin{pmatrix} 1 & b_{12} & \cdots & b_{1m} \\ 0 & 1 & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{m \times m} \\
=:&ED_uB_u \\
=:&EB
\end{aligned}
$$

where diagonal entries of D_u are $d_{ii} = |u_i| > 0$, and diagonal entries of B_u are $b_{ii} = 1$, $1 \leq i \leq m$. Therefore, the upper triangular matrix $B = D_uB_u$ has non-zero diagonal entries.

The hat matrix H is

$$
\begin{aligned}
H =&X_t[X_t^T X_t]^{-1} X_t^T, \text{ where } X_t = EA \\
=&EA[A^T E^T EA]^{-1} A^T E^T, \text{ where } E^T E = I_m \\
=&EA[A^T A]^{-1} A^T E^T \\
=&E \begin{pmatrix} I_{p+1} & 0_{(p+1)\times(m-p-1)} \\ 0_{(m-p-1)\times(p+1)} & 0_{m-p-1} \end{pmatrix} E^T
\end{aligned}
$$

The hat matrix H_1 is

$$
\begin{aligned}
H_1 =&X_y[X_y^T X_y]^{-1} X_y^T, \text{ where } X_y = EB \\
=&EB[B^T E^T EB]^{-1} B^T E^T, \text{ where } E^T E = I_m \\
=&EB[B^T B]^{-1} B^T E^T \\
=&EI_m E^T
\end{aligned}
$$

Therefore,

$$\begin{aligned}
HH_1 =& E \begin{pmatrix} I_{p+1} & 0_{(p+1)\times(m-p-1)} \\ 0_{(m-p-1)\times(p+1)} & 0_{m-p-1} \end{pmatrix} E^T E I_m E^T, \text{ where } E^T E = I_m \\
=& E \begin{pmatrix} I_{p+1} & 0_{(p+1)\times(m-p-1)} \\ 0_{(m-p-1)\times(p+1)} & 0_{m-p-1} \end{pmatrix} I_m E^T \\
=& E \begin{pmatrix} I_{p+1} & 0_{(p+1)\times(m-p-1)} \\ 0_{(m-p-1)\times(p+1)} & 0_{m-p-1} \end{pmatrix} E^T \\
=& H
\end{aligned}$$

More General Version of Theorem 3. *In the uni-dimensional case for the instrument $\vec{Z}$, if models in Stage One and Stage Two are general additive models that have the following forms:*

$$\hat{t} = \omega_1 \text{feature}_1(\vec{X}) + \cdots + \omega_{k_1} \text{feature}_{k_1}(\vec{X}) + \omega_{k_1+1} \text{feature}_{k_1+1}(\vec{Z})$$

where $\text{feature}_j(\vec{X})$, $j = 1, \cdots, k_1$ *is a linear or non-linear function of* $\vec{X}$ *and* $\text{feature}_{k_1+1}(\vec{Z})$ *is a linear or non-linear function of* $\vec{Z}$.

$$\begin{aligned}
\hat{y} =& \beta_1 \text{feature}_1(\vec{X}) + \cdots + \beta_{k_1} \text{feature}_{k_1}(\vec{X}) + \beta_{k_1+1} \hat{t} + \\
& \beta_{k_1+2} \text{feature}_{k_1+1}(\vec{X}, \hat{t}) + \cdots + \beta_{k_1+k_2+1} \text{feature}_{k_1+k_2}(\vec{X}, \hat{t})
\end{aligned}$$

where $\text{feature}_j(\vec{X})$, $j = 1, \cdots, k_1$ *is a linear or non-linear function of* $\vec{X}$ *and* $\text{feature}_j(\vec{X}, \hat{t})$, $j = k_1 + 1, \cdots, k_1 + k_2$ *is a non-linear function of* $(\vec{X}, \hat{t})$ *or a non-linear function of only* $\hat{t}$.

In other word, the following conditions are satisfied:

1. the models in Stage One and Stage Two are general additive models that satisfied all the conditions in **Theorem 2.**

2. the model in Stage Two contains non-linear features of the predicted values of the treatment $\hat{t}$ and the covariates x, which can be non-linear features of only x, non-linear features of only $\hat{t}$ and the interaction terms of x and $\hat{t}$.

then the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H \vec{y} + \vec{y}^T H \vec{y} \leq \epsilon'$) is also never active. This statement is true regardless of the dimension of $\vec{X}$.

Proof. Recall that **Theorem 2.** holds because of $H = H_1$. In the **More General Version of Theorem 3.**, although $H \neq H_1$, $H - HH_1 = 0$.

In Stage Two of our method, the optimal / minimum solution of the objective function (2) ($\beta^T X_y^T X_y \beta - 2\beta^T X_y^T \vec{y} + \vec{y}^T \vec{y}$) is $\hat{\beta}_{min} = [X_y^T X_y]^{-1} X_y^T \vec{y}$ and $\hat{y} = X_y \hat{\beta}_{min} = X_y [X_y^T X_y]^{-1} X_y^T \vec{y} = H_1 \vec{y}$.

The objective function on the left side of the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H \vec{y} + \vec{y}^T H \vec{y}$) is:

$$\begin{aligned}
&||H(\vec{y} - \hat{y})||^2 \\
=&||H\vec{y} - H\hat{y})||^2 \\
=&||H\vec{y} - HH_1 y)||^2 \text{ where } \hat{y} = H_1 y \\
=&||(H - HH_1)y)||^2 \text{ where } H - HH_1 = 0 \\
=&0 \leq \epsilon' \text{ where } \epsilon' > 0
\end{aligned}$$

Therefore, the constraint (3) ($\beta^T X_y^T H X_y \beta - 2\beta^T X_y^T H \vec{y} + \vec{y}^T H \vec{y} \leq \epsilon'$) is always satisfied (or never active).

Next prove $H - HH_1 = 0$

Recall the notation of the models in Stage One and Stage Two in the **More General Version of Theorem 3.** and define the predictor matrices as follows:

Define the predictor matrix of t in Stage One:

$$X_t = (\text{feature}_1(\vec{X}), \cdots, \text{feature}_{k_1}(\vec{X}), \text{feature}_{k_1+1}(\vec{Z}))_{n \times (k_1+1)} \in \mathbb{R}^{n \times (k_1+1)}$$

Define the predictor matrix of y in Stage Two:

$$X_y = (\text{feature}_1(\vec{X}), \cdots, \text{feature}_{k_1}(\vec{X}), \hat{t}, \text{feature}_{k_1+1}(\vec{X}, \hat{t}), \cdots, \text{feature}_{k_1+k_2}(\vec{X}, \hat{t}))_{n \times (k_1+k_2+1)} \in \mathbb{R}^{n \times m}$$

where $m = k_1 + k_2 + 1$.

Next Use the Gram-Schmidt algorithm to construct an orthogonal set of unit vectors.

Step 1: $u_1 = \text{feature}_1(\vec{X})$, $e_1 = \frac{u_1}{|u_1|}$

Step 2: $u_2 = \text{feature}_2(\vec{X}) - \frac{\text{feature}_2(\vec{X})\cdot u_1}{|u_1|^2} u_1$, $e_2 = \frac{u_2}{|u_2|}$

$\vdots$

Step k_1+1: $u_{k_1+1} = \hat{t} - \frac{\hat{t}\cdot u_1}{|u_1|^2}u_1 - \cdots - \frac{\hat{t}\cdot u_{k_1}}{|u_{k_1}|^2}u_{k_1}$, $e_{k_1+1} = \frac{u_{k_1+1}}{|u_{k_1+1}|}$

Step k_1+2: $u_{k_1+2} = \text{feature}_{k_1+1}(\vec{X},\hat{t}) - \frac{\text{feature}_{k_1+1}(\vec{X},\hat{t})\cdot u_1}{|u_1|^2}u_1 - \cdots - \frac{\text{feature}_{k_1+1}(\vec{X},\hat{t})\cdot u_{k_1+1}}{|u_{k_1+1}|^2}u_{k_1+1}$, $e_{k_1+2} = \frac{u_{k_1+2}}{|u_{k_1+2}|}$

$\vdots$

Step m: $u_m = \text{feature}_{k_1+k_2}(\vec{X},\hat{t}) - \frac{\text{feature}_{k_1+k_2}(\vec{X},\hat{t})\cdot u_1}{|u_1|^2}u_1 - \cdots - \frac{\text{feature}_{k_1+k_2}(\vec{X},\hat{t})\cdot u_{m-1}}{|u_{m-1}|^2}u_{m-1}$, $e_m = \frac{u_m}{|u_m|}$

Therefore the predictor matrix and t and y can be written as follows,

$$
\begin{aligned}
X_t =& (\text{feature}_1(\vec{X}), \cdots, \text{feature}_{k_1}(\vec{X}), \text{feature}_{k_1+1}(\vec{Z})) \\
=& (u_1, u_2, \cdots, u_m)
\begin{pmatrix}
1 & a_{12} & \cdots & a_{1(k_1+1)} \\
0 & 1 & \cdots & a_{2(k_1+1)} \\
\vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 1/\omega_{k_1+1} \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 0
\end{pmatrix}_{m\times(k_1+1)} \\
=:& UA_u = (e_1, e_2, \cdots, e_m)
\begin{pmatrix}
|u_1| & 0 & \cdots & 0 \\
0 & |u_2| & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & |u_m|
\end{pmatrix}_{m\times m}
\begin{pmatrix}
1 & a_{12} & \cdots & a_{1(k_1+1)} \\
0 & 1 & \cdots & a_{2(k_1+1)} \\
\vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 1/\omega_{k_1+1} \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 0
\end{pmatrix}_{m\times(k_1+1)} \\
=:& ED_uA_u \\
=:& EA
\end{aligned}
$$

where diagonal entries of D_u are $d_{ii} = |u_i| > 0$, and diagonal entries of A_u are $a_{ii} = 1$, $1 \leq i \leq k_1$ and $a_{(k_1+1)(k_1+1)} = 1/\omega_{k_1+1}$. Therefore, the upper trapezoidal matrix $A = D_u A_u$ has non-zero diagonal entries.

$$
\begin{aligned}
X_y =& (\text{feature}_1(\vec{X}), \cdots, \text{feature}_{k_1}(\vec{X}), \hat{t}, \text{feature}_{k_1+1}(\vec{X}, \hat{t}), \cdots, \text{feature}_{k_1+k_2}(\vec{X}, \hat{t})) \\
=& (u_1, u_2, \cdots, u_m) \begin{pmatrix} 1 & b_{12} & \cdots & b_{1m} \\ 0 & 1 & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{m \times m} \\
=:& UB_u = (e_1, e_2, \cdots, e_m) \begin{pmatrix} |u_1| & 0 & \cdots & 0 \\ 0 & |u_2| & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & |u_m| \end{pmatrix}_{m \times m} \begin{pmatrix} 1 & b_{12} & \cdots & b_{1m} \\ 0 & 1 & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{m \times m} \\
=:& ED_u B_u \\
=:& EB
\end{aligned}
$$

where diagonal entries of D_u are $d_{ii} = |u_i| > 0$, and diagonal entries of B_u are $b_{ii} = 1$, $1 \leq i \leq m$. Therefore, the upper triangular matrix $B = D_u B_u$ has non-zero diagonal entries.

The hat matrix H is

$$
\begin{aligned}
H =& X_t [X_t^T X_t]^{-1} X_t^T, \text{ where } X_t = EA \\
=& EA[A^T E^T EA]^{-1} A^T E^T, \text{ where } E^T E = I_m \\
=& EA[A^T A]^{-1} A^T E^T \\
=& E \begin{pmatrix} I_{k_1+1} & 0_{(k_1+1) \times k_2} \\ 0_{k_2 \times (k_1+1)} & 0_{k_2} \end{pmatrix} E^T
\end{aligned}
$$

The hat matrix H_1 is

$$\begin{aligned} H_1 =& X_y[X_y^T X_y]^{-1} X_y^T, \text{ where } X_y = EB \\ =& EB[B^T E^T EB]^{-1} B^T E^T, \text{ where } E^T E = I_m \\ =& EB[B^T B]^{-1} B^T E^T \\ =& EI_m E^T \end{aligned}$$

Therefore,

$$\begin{aligned} HH_1 =& E \begin{pmatrix} I_{k_1+1} & 0_{(k_1+1)\times k_2} \\ 0_{k_2\times(k_1+1)} & 0_{k_2} \end{pmatrix} E^T E I_m E^T, \text{ where } E^T E = I_m \\ =& E \begin{pmatrix} I_{k_1+1} & 0_{(k_1+1)\times k_2} \\ 0_{k_2\times(k_1+1)} & 0_{k_2} \end{pmatrix} I_m E^T \\ =& E \begin{pmatrix} I_{k_1+1} & 0_{(k_1+1)\times k_2} \\ 0_{k_2\times(k_1+1)} & 0_{k_2} \end{pmatrix} E^T \\ =& H \end{aligned}$$

Conclusion 1. *In the uni-dimensional case for the instrument $\vec{Z}$, if the models in Stage One and Stage Two are both linear models, our two stage method is identical with the traditional two stage.*

Proof. Since models in both stages of the traditional two stage method are linear models, according to Theorem 2, the constraint is never active. Therefore, the traditional two stage method and our two stage method return the same results.

Conclusion 2. *In the multi-dimensional case for the instrument $\vec{Z}$, if the models in Stage One and Stage Two are both linear models, our two stage method cannot find a better optimization / minimization solution than the traditional two stage method, but can at least identify when the model is bad.*

Proof. Since models in both stages of the traditional two stage method are linear models, according to Theorem 1, the constraint is either not active or cannot be satisfied. More specifically, once the constraint is active, there is no solution for the constrained Stage

Two. Therefore, the traditional two stage method and our two stage method return the same results.

Conclusion 3. *If the models in Stage One and Stage Two are both linear models, our two stage method cannot fix a bad modelling choice i.e. our two stage method cannot improve the model performance by adjusting coefficients under the linear framework.*

Chapter 7

Pseudo Code

In order to generalize the square loss, we replace the inner product in the square loss function with a kernel function, which generalize the least square solution to the kernel least square solution. The pseudo codes of the kernel least square methods are as follows. Note that we only present the general procedures of the kernel least square methods but do not provide the closed form solution. When there is not a closed form for the optimization problem, different techniques are applied in practice to reach a local optimal solution.

7.1 Two Stage Method

Stage One

1. construct the predictor matrix of t in Stage One, $X^{(t)}$, where matrix $X^{(t)}$ is defined element-wise by $x_{ij}^{(t)} = function_j(x_{i\cdot}, z_{i\cdot})$;

2. compute r_1^* by minimizing the square loss: $r_1^* \in \text{argmin}_{r_1} loss(r_1) = ||t - X^{(t)}\omega_{r_1}||_2^2$, where $\omega_{r_1} = {X^{(t)}}^T r_1 = \sum_{i=1}^{n} {x^{(t)}}_{i\cdot}^T r_{1i}$, and obtain the least square solution of $\hat{\omega}_{r_1}$, $\hat{\omega}_{r_1} = \sum_{i=1}^{n} {x^{(t)}}_{i\cdot}^T r_{1i}^*$;

3. obtain the least square solution of the treatment $\hat{t}$, $\hat{t}_{i\cdot} = {x^{(t)}}_{i\cdot}\hat{\omega}_{r_1} = {x^{(t)}}_{i\cdot} \sum_{i=1}^{n} {x^{(t)}}_{i\cdot}^T r_{1i}^* = \sum_{i=1}^{n} {x^{(t)}}_{i\cdot} {x^{(t)}}_{i\cdot}^T r_{1i}^*$;

4. if desired, replace the inner product with a kernel function $k(\cdot,\cdot)$, and obtain the kernel least square solution of $\hat{t}$, $\hat{t}_{i\cdot} = \sum_{i=1}^{n} k({x^{(t)}}_{i\cdot}, {x^{(t)}}_{i\cdot}) r_{1i}^*$.

Stage Two

1. construct the predictor matrix of y in Stage One, $X^{(y)}$, where matrix $X^{(y)}$ is defined element-wise by $x_{ij}^{(y)} = function_j(x_{i\cdot}, \hat{t}_i)$;

2.compute r_2^* by minimizing the square loss: $r_2^* \in \text{argmin}_{r_2} loss(r_2) = ||y - X^{(y)}\beta_{r_2}||_2^2 + C(||H\vec{y} - HX^{(y)}\beta_{r_2}||_2^2 - \epsilon)$, where $H = X_t(X_t^T X_t)^{-1}X_t^T$ and $\beta_{r_2} = X^{(y)^T} r_2 = \sum_{i=1}^{n} x^{(y)^T}_{i\cdot} r_{2i}$, and obtain the least square solution of $\hat{\beta}_{r_2}$, $\hat{\beta}_{r_2} = \sum_{i=1}^{n} x^{(y)^T}_{i\cdot} r^*_{2i}$;

3. obtain the least square solution of the outcomes $\hat{y}$, $\hat{y}_{i\cdot} = x^{(y)}{}_{i\cdot}\hat{\beta}_{r_2} = x^{(y)}{}_{i\cdot} \sum_{i=1}^{n} x^{(y)^T}_{i\cdot} r^*_{2i} = \sum_{i=1}^{n} x^{(y)}{}_{i\cdot} x^{(y)^T}_{i\cdot} r^*_{2i}$;

4. if desired, replace the inner product with a kernel function $k(\cdot,\cdot)$, and obtain the kernel least square solution of $\hat{y}$, $\hat{y}_{i\cdot} = \sum_{i=1}^{n} k(x^{(y)}{}_{i\cdot}, x^{(y)}{}_{i\cdot}) r^*_{2i}$.

7.2 One Stage Method

1. construct the predictor matrix of t in Stage One, $X^{(t)}$, where matrix $X^{(t)}$ is defined element-wise by $x^{(t)}_{ij} = function_j(x_{i\cdot}, z_{i\cdot})$ and the predictor matrix of y in Stage One, $X^{(y)}$, where matrix $X^{(y)}$ is defined element-wise by $x^{(y)}_{ij} = function_j(x_{i\cdot}, \hat{t}_i)$;

2. compute r_1^* and r_2^* by minimizing the square loss: $loss(r_1, r_2) = ||y - X^{(y)}(r_1)\beta_{r_2}||^2 + C_1(||t - X^{(t)}\omega_{r_1}||^2 - \epsilon_1) + C_2(||H\vec{y} - HX^{(y)}(r_1)\beta_{r_2}||^2 - \epsilon_2)$, where $H = X_t(X_t^T X_t)^{-1}X_t^T$, $X^{(y)}(r_1) = function(\vec{X}, \hat{t}(r_1))$, $\hat{t}(r_1) = X^{(t)^T}\omega_{r_1}$, $\omega_{r_1} = X^{(t)^T} r_1 = \sum_{i=1}^{n} x^{(t)^T}_{i\cdot} r_{1i}$ and $\beta_{r_2} = X^{(y)^T} r_2 = \sum_{i=1}^{n} x^{(y)^T}_{i\cdot} r_{2i}$, and obtain the least square solutions of $\hat{\omega}_{r_1}$, $\hat{\omega}_{r_1} = \sum_{i=1}^{n} x^{(t)^T}_{i\cdot} r^*_{1i}$ and $\hat{\beta}_{r_2}$, $\hat{\beta}_{r_2} = \sum_{i=1}^{n} x^{(y)^T}_{i\cdot} r^*_{2i}$;

3. obtain the least square solutions of the treatment $\hat{t}$, $\hat{t}_{i\cdot} = x^{(t)}{}_{i\cdot}\hat{\omega}_{r_1} = x^{(t)}{}_{i\cdot} \sum_{i=1}^{n} x^{(t)^T}_{i\cdot} r^*_{1i} = \sum_{i=1}^{n} x^{(t)}{}_{i\cdot} x^{(t)^T}_{i\cdot} r^*_{1i}$ and the outcomes $\hat{y}$, $\hat{y}_{i\cdot} = x^{(y)}{}_{i\cdot}\hat{\beta}_{r_2} = x^{(y)}{}_{i\cdot} \sum_{i=1}^{n} x^{(y)^T}_{i\cdot} r^*_{2i} = \sum_{i=1}^{n} x^{(y)}{}_{i\cdot} x^{(y)^T}_{i\cdot} r^*_{2i}$;

4. if desired, replace the inner product with a kernel function $k(\cdot,\cdot)$, and obtain the kernel least square solutions of $\hat{t}$, $\hat{t}_{i\cdot} = \sum_{i=1}^{n} k(x^{(t)}{}_{i\cdot}, x^{(t)}{}_{i\cdot}) r^*_{1i}$ and $\hat{y}$, $\hat{y}_{i\cdot} = \sum_{i=1}^{n} k(x^{(y)}{}_{i\cdot}, x^{(y)}{}_{i\cdot}) r^*_{2i}$.

Chapter 8

Simulation

8.1 Simulation Results that Accord with Theorems

8.1.1 Our Two Stage Method

1-D Instrument

This subsection presents the simulations that accord with **Theorem 2** and **Theorem 3** In the uni-dimensional case for the instrument $\vec{Z}$, if the prediction model in the first stage is a linear model and the prediction model in the second stage is a linear or general additive model, the constraint is never active.

Table 8.1: 1-D Instrument: Results using Our Two-Stage Methods

Index	**Data Generation**	**Error Term**	**Prediction Model**	**Fraction of time that the constraint holds / Fraction of time that the left side of the constraint is zero**
Row 1-3 state that the constraint is always satisfied (or never active) regardless of the type of the error term when models in both stages are linear models and the data generation and prediction models share the same model forms.				
1	$t = x + z + e_1$ $y = x + t + e_2$	Gaussian error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t}$	100% / 100%
2	$t = x + z + e_1$ $y = x + t + e_2$	mixture error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t}$	100% / 100%
3	$t = x + z + e_1$ $y = x + t + e_2$	fanning error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t}$	100% / 100%

Row 4-6 state that the constraint is always satisfied (or never active) regardless of the type of the error term when the model in the first stage is linear, the model in the second stage is a general additive model, and the data generation and prediction models share the same model forms.				
4	$t = x + z + e_1$ $y = x + t + t^2 + e_2$	Gaussian error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t} + \hat{t}^2$	100% / 100%
5	$t = x + z + e_1$ $y = x + t + t^2 + e_2$	mixture error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t} + \hat{t}^2$	100% / 100%
6	$t = x + z + e_1$ $y = x + t + t^2 + e_2$	fanning error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t} + \hat{t}^2$	100% / 100%
Row 7-10 state that the constraint is always satisfied (or never active) even when the data generation and prediction models do not have the same model forms.				
7	$t = x^2 + z^2 + e_1$ $y = x^2 + t^2 + e_2$	fanning error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t}$	100% / 100%
8	$t = sin(x) + cos(z) + e_1$ $y = cos(x) + sin(t) + e_2$	fanning error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t}$	100% / 100%
9	$t = x^2 + z^2 + e_1$ $y = x^2 + t^2 + e_2$	fanning error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t} + \hat{t}^2 + x\hat{t}$	100% / 100%
10	$t = sin(x) + cos(z) + e_1$ $y = cos(x) + sin(t) + e_2$	fanning error	$\hat{t} = x + z$ $\hat{y} = x + \hat{t} + \hat{t}^2 + x\hat{t}$	100% / 100%

Row 1-3 of Table 8.1 show that **Theorem 2** is true regardless of the type of the error term.

Row 4-6 of Table 8.1 show that **Theorem 3** is true regardless of the type of the error term.

Row 7-10 of Table 8.1 show that **Theorem 2** and **Theorem 3** hold as long as the prediction models satisfy their conditions, regardless of the forms of the data generation functions.

Note that the **More General Version of Theorem 2** and the **More General Version of Theorem 3** can also be verified by similar simulations.

2^+-D Instrument

This subsection presents the simulations that accord with **Theorem 1**. If the prediction model in the first stage is a linear or general additive model and the prediction model in the second stage is a linear model, either the constraint is not active, or there is no feasible solution.

Table 8.2: 2+-D Instrument: Results using Our Two-Stage Methods.

Index	**Data Generation**	**Prediction Model**	**Epsilon Percent** γ	**Fraction of time that the constraint holds**	**Fraction of time that there is no feasible solution**	**Total Fraction**
Row 1-3 state that in the multi-dimensional cases for instrument z, either the constraint is not active or there is no feasible solution when models in both stage are linear and the data generation and prediction models share the same model forms.						
1	$t = x + z_1 + z_2 + e_1$ $y = x + t + e_2$	$\hat{t} = x + z_1 + z_2$ $\hat{y} = x + \hat{t}$	0.5%	99.9%	0.1%	100%
2	$t = x + z_1 + z_2 + e_1$ $y = x + t + e_2$	$\hat{t} = x + z_1 + z_2$ $\hat{y} = x + \hat{t}$	0.1%	84%	16%	100%
3	$t = x_1 + x_2 + z_1 + z_2 + z_3 + e_1$ $y = x_1 + x_2 + t + e_2$	$\hat{t} = x_1 + x_2 + z_1 + z_2 + z_3$ $\hat{y} = x_1 + x_2 + \hat{t}$	0.5%	99.5%	0.5%	100%
Row 4-6 state that in the multi-dimensional cases for instrument z, either the constraint is not active or there is no feasible solution when the model in the first stage is a general additive model, the model in the second stage is linear and the data generation and prediction models share the same model forms.						
4	$t = x_1 + x_2 + z_1 + z_2 + z_3 + e_1$ $y = x_1 + x_2 + t + e_2$	$\hat{t} = x_1 + x_2 + z_1 + z_2 + z_3$ $\hat{y} = x_1 + x_2 + \hat{t}$	0.1%	64.8%	35.2%	100%
5	$t = x_1 + x_2 + z_1 + z_2 + x_1^2 + x_2^2 + z_1^2 + z_2^2 + e_1$ $y = x_1 + x_2 + t + e_2$	$\hat{t} = x_1 + x_2 + z_1 + z_2 + x_1^2 + x_2^2 + z_1^2 + z_2^2$ $\hat{y} = x_1 + x_2 + \hat{t}$	0.5%	84.8%	15.2%	100%
6	$t = x_1 + x_2 + z_1 + z_2 + x_1 z_1 + x_2 z_2 + e_1$ $y = x_1 + x_2 + t + e_2$	$\hat{t} = x_1 + x_2 + z_1 + z_2 + x_1 z_1 + x_2 z_2$ $\hat{y} = x_1 + x_2 + \hat{t}$	0.5%	96.0%	4.0%	100%
Row 7-10 state that in the multi-dimensional cases for instrument z, either the constraint is not active or there is no feasible solution even when the data generation and prediction models do not have the same model forms.						

7	$t = x_1^2+x_2^2+z_1^2+z_2^2+e_1$ $y = x_1^2 + x_2^2 + t^2 + e_2$	$\hat{t} = x_1 + x_2 + z_1 + z_2$ $\hat{y} = x_1 + x_2 + \hat{t}$	0.1%	61.6%	38.4%	100%
8	$t = sin(x_1) + sin(x_2) +$ $cos(z_1) + cos(z_2) + e_1$ $y = sin(x_1) + sin(x_2) +$ $cos(t) + e_2$	$\hat{t} = x_1 + x_2 + z_1 + z_2$ $\hat{y} = x_1 + x_2 + \hat{t}$	0.1%	71.7%	28.3%	100%
9	$t = x_1^2+x_2^2+z_1^2+z_2^2+e_1$ $y = x_1^2 + x_2^2 + t^2 + e_2$	$\hat{t} = x_1+x_2+z_1+z_2+$ $x_1z_1 + x_2z_2$ $\hat{y} = x_1 + x_2 + \hat{t}$	0.5%	97.8%	2.0%	100%
10	$t = sin(x_1) + sin(x_2) +$ $cos(z_1) + cos(z_2) + e_1$ $y = sin(x_1) + sin(x_2) +$ $cos(t) + e_2$	$\hat{t} = x_1+x_2+z_1+z_2+$ $x_1z_1 + x_2z_2$ $\hat{y} = x_1 + x_2 + \hat{t}$	0.5%	90.9%	9.1%	100%

In Table 8.2, the sum of percentages of the fifth and sixth columns is always 100%.

Row 1-6 of Table 8.2 shows that **Theorem 1** is true when the prediction model is the same as the generation model. These are similar experiments to Rows 1-6 of Table 8.1, but for the multi-dimensional case.

Row 7-10 of Table 8.2 shows that **Theorem 1** holds as long as the prediction models satisfy their conditions, regardless of the forms of the data generation functions.

Note that the **More General Version of Theorem 1** can also be verified by similar simulations.

8.1.2 Our One Stage Method

The one-stage method is not limited by the restrictions given within **Theorems 1-3** and **More General Version of Theorem 1-3**. In particular, the one-stage procedure can provide more flexibility in the optimization process. In this section, we demonstrate this advantage of the one-stage method. This section compares the simulation results of our two stage method and our one stage method. The results show that the theorems in Chapter 6 only hold for our two stage method but are not true for our one stage method.

Table 8.3: Results using Our One-Stage Method.

Index	**Data Generation**	**Prediction Model**	**Fraction of time that the constraint holds**	**Fraction of time that there is a feasible solution for the two-stage method**	**Fraction of time that there is a feasible solution for the one-stage method**	**Fraction of time that there is no feasible solution**	**Total Fraction**
Row 1-4 state that the fractions of time that there is a feasible solution are the same for our two-stage and one-stage method when models in both stages are general additive models.							
1	$t = x+z+xz+e_1$ $y = x+t+xt+e_2$	$\hat{t} = x+z+xz$ $\hat{y} = x+\hat{t}+x\hat{t}$	99.0%	1.0%	1.0%	0.0%	100%
2	$t = x+z+z^2+e_1$ $y = x+t+t^2+e_2$	$\hat{t} = x+z+z^2$ $\hat{y} = x+\hat{t}+\hat{t}^2$	97.0%	3.0%	3.0%	0.0%	100%
3	$t = x+x^2+z+$ z^2+e_1 $y = x+x^2+t+$ t^2+e_2	$\hat{t} = x+z+z^2$ $\hat{y} = x+\hat{t}+\hat{t}^2$	96.0%	4.0%	4.0%	0.0%	100%
4	$t = x+z_1+z_2+$ $z_1^2+z_2^2+e_1$ $y = x+t+t^2+e_2$	$\hat{t} = x+z_1+z_2+$ $z_1^2+z_2^2$ $\hat{y} = x+\hat{t}+\hat{t}^2$	98.0%	1.0%	1.0%	1.0%	100%
Row 5-8 state that the fraction of time that there is a feasible solution for our one-stage method are greater than that for two-stage method when the model in the first stage is a linear or general additive model and the model in the second stage is linear.							
5	$t = x+z+xz+e_1$ $y = x+t+e_2$	$\hat{t} = x+z+xz$ $\hat{y} = x+\hat{t}$	92.0%	0.0%	8.0%	0.0%	100%
6	$t = x+z+z^2+e_1$ $y = x+t+e_2$	$\hat{t} = x+z+z^2$ $\hat{y} = x+\hat{t}$	96.0%	0.0%	4.0%	0.0%	100%
7	$t = x+x^2+z+$ z^2+e_1 $y = x+t+e_2$	$\hat{t} = x+z+xz$ $\hat{y} = x+\hat{t}$	88.0%	0.0%	12.0%	0.0%	100%
8	$t = x+z_1+z_2+e_1$ $y = x+t+e_2$	$\hat{t} = x+z_1+z_2$ $\hat{y} = x+\hat{t}$	96%	0.0%	2.0%	2.0%	100%

In Table 8.3, the sum of percentages, which is always 100%, is the fourth and second last columns, plus the maximum of the fifth and sixth columns.

Row 1-4 of Table 8.3 show that our general two-stage method and one-stage method perform equally well when both of the two models are general additive models.

Row 5-8 of Table 8.3 show that our one-stage method is not subject to **Theorem 1** and can outperform our general two-stage method. The two-stage method cannot provide a feasible solution in some cases where the one-stage method has a feasible solution.

Note that **Theorem 2** and **Theorem 3** also do not hold for our one-stage method.

8.2 Identify Invalid Instruments

In the simulations below, we always assume that the models in both stage are true prediction models; in other words, we assumed we had made good modeling choices. In what follows, we assume we correctly specified the model form in the first stage, but that there are unknown covariates in the second stage. We construct 1000 simulations for both the valid instrument case and the invalid instrument case respectively. We use the following data generation mechanism.

Data Generation (Construct Valid Instrument):

$$\begin{aligned} t &= x + z + z^2 + e_1 \\ y &= x + t + e_2. \end{aligned}$$

Data Generation (Construct Invalid Instrument):

$$\begin{aligned} x_{UN} &= z^3 \\ t &= x + z + z^2 + e_1 \\ y &= x + x_{UN} + t + e_2. \end{aligned}$$

where $x \sim N(0,1)$, $z \sim N(0,1)$ and $\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \sim N\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \Big)$.

Table 8.4: Confusion Matrix by Checking Validity after Modelling ($\gamma = 10\%$)

n = 2000	The instrument is predicted to be invalid (The constraint is not satisfied)	The instrument is predicted to be valid (The constraint is not active)
The instrument is invalid	1000	0
The instrument is valid	0	1000

Table 8.5: Confusion Matrix by Our Two Stage Method ($\gamma = 10\%$)

n = 2000	The instrument is predicted to be invalid (There is no feasible solution)	The instrument is predicted to be valid (There is a feasible solution)
The instrument is invalid	1000	0
The instrument is valid	0	1000

Here, the predictive model is the same form, but the unknown covariates make the instrument invalid.

Prediction Model:

$$\begin{aligned} \hat{t} &= x + z + z^2 \\ \hat{y} &= x + \hat{t}. \end{aligned}$$

We show that our two stage method can identify when the instrument is valid and when it is not. To do this, we ran 2000 simulations, where 1000 of them used a valid instrument, and 1000 of them used an invalid instrument. We used our two stage method in two ways: first, we did not enforce the validity constraint and instead, checked whether the validity constraint held afterwards, which is in Table 8.4. Second, in Table 8.5, we enforced the validity constraint directly. These tables show that our two-stage method identified whether the instrument is valid in each of the simulations, even without the constraint enforced, which also meant that the instrument was still valid when the constraint was enforced.

8.3 Stability of Our Method

This section shows that when the instrument is valid, the coefficients obtained by our new method are more accurate than that obtained by the traditional 2SLS method. In Figure 8.1, we use the coefficient π to quantify the strength of the instrument and the absolute value of the bias of the median estimate with respect to the coefficient β in the second stage as the measurement of estimation accuracy.

$$\begin{aligned} t &= x + \pi(z + z^2) + e_1 \\ y &= x + \beta t + e_2 \end{aligned}$$

where $\beta = 1$, $x \sim N(0,1)$, $z \sim N(0,1)$ and $\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right)$.

According to Figure 8.1, the decreasing lines show that as the first stage become stronger, the estimation of the true causal effect of the treatment t on outcomes y becomes more accurate. The fact that the blue line is higher than the orange line and the green line shows that our two stage method outperforms the traditional 2SLS method with more accurate estimation of the true causal effect.

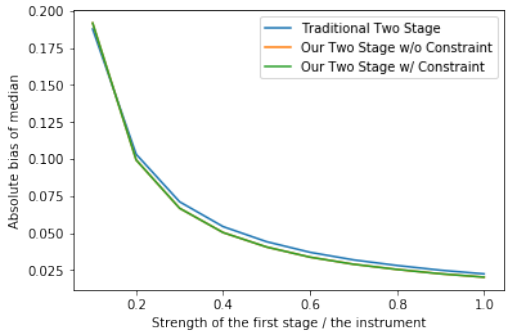

Figure 8.1: The (The orange line overlaps with the green line.)

8.4 Detect and Fix Bad Modeling Choices

In the following example, we have a good instrument but a poor modeling choice, in that the data generation is a quadratic model, but the prediction model is linear. If we use a quadratic prediction model for the remainder r, we can detect that the modeling choice was poor. In that case, we can fix the modeling choice for the treatment $\hat{t}$ so that the instrument appears to be valid.

In particular, we use the following data generation process to construct a valid instrument case:

$$\begin{aligned} t &= x + z + z^2 + e_1 \\ y &= x + t + e_2. \end{aligned}$$

where $x \sim N(0,1)$, $z \sim N(0,1)$ and $\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \sim N\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \Big)$.

The form of predictive models for $\hat{t}$ and $\hat{y}$ are:

$$\begin{aligned} \hat{t} &= x + z \\ \hat{y} &= x + \hat{t}. \end{aligned}$$

Here are the predictor matrices used in the three prediction models for the treatment t, outcomes y and the remainder r respectively:

$$\begin{aligned} Xt &= (x, z) \\ Xy &= (x, \hat{t}) \\ Xr &= (x, z, z^2). \end{aligned}$$

Here we use a more flexible class of prediction models for the remainder r than that for the treatment t to detect a bad modelling choice while the instrument is actually valid. In this case, the quadratic model form for the remainder r can detect the quadratic dependence on the instrument z which cannot be found out by the linear model for the treatment t in the first stage. Furthermore, since the remainder can be well-predicted by a more flexible

model class, the loss between r and its estimated value $\hat{r}$ is significantly smaller than the loss between r and zero. Therefore, the constraint is not satisfied due to the bad modelling choice. However, if we then apply the quadratic model form in the first stage for estimation of the treatment t, the constraint cannot be satisfied again and the bad modelling choice is fixed. The predictor matrices used in fixed modelling process for the treatment t, outcomes y and the remainder r respectively:

$$\begin{aligned} Xt &= (x, z, z^2) \\ Xy &= (x, \hat{t}) \\ Xr &= (x, z, z^2, z^3). \end{aligned}$$

The values of loss functions on both sides of the constraint, $loss(r, \hat{r})$ and $loss(r, 0)$ for two modelling choices are as follows.

Table 8.6: Values of loss on both sides on the constraint

	$loss(r, 0)$	$loss(r, \hat{r})$	True epsilon percentage: $loss(r, \hat{r})/loss(r, 0)$
Linear first stage: $Xt = (x, z)$	44.4444	34.5046	77.64%
Quadratic first stage: $Xt = (x, z, z^2)$	34.5091	34.5028	99.98%

The results in table 8.6 show that if we choose a epsilon percentage γ equals to 10%, a bad modelling choice, the linear first stage will be rejected while the fixed quadratic first stage will be accepted.

According to the example above, we demonstrate than a bad modelling choice can be detected and fixed, the following example shows that a bad / invalid instrument cannot be fixed.

We continue to play with the invalid instrument case in **Section 8.2**. In **Section 8.2**, we use quadratic model form in the first stage to model the relationship between z and t. The predictor matrices used in three models for the treatment t, outcomes y and the

remainder r respectively:

$$\begin{aligned} Xt &= (x, z, z^2) \\ Xy &= (x, \hat{t}) \\ Xr &= (x, z, z^2, z^3). \end{aligned}$$

In order model the cubic dependence of outcomes y on the instrument z through the unknown covariates $x_{UN} - z^3$, we then apply the cubic model in the first stage.

The predictor matrices used for new modelling choices of the treatment t, outcomes y and the remainder r respectively:

$$\begin{aligned} Xt &= (x, z, z^2, z^3) \\ Xy &= (x, \hat{t}) \\ Xr &= (x, z, z^2, z^3, z^4). \end{aligned}$$

The values of loss functions on both sides of the constraint, $loss(r, \hat{r})$ and $loss(r, 0)$ for two modelling choices are as follows.

Table 8.7: Values of loss on both sides on the constraint

	$loss(r, 0)$	$loss(r, \hat{r})$	True epsilon percentage: $loss(r, \hat{r})/loss(r, 0)$
Linear first stage: $Xt = (x, z, z^2)$	42.9038	35.4535	82.63%
Quadratic first stage: $Xt = (x, z, z^2, z^3)$	42.9131	35.3490	82.64%

The results in table 8.7 show that the true epsilon percentages before and after trying to fix a bad instrument are very close to each other and cannot be told apart by a predetermined epsilon percentage γ.

Chapter 9

Real Data

In this chapter, we tested our two-stage and one-stage methods on one real world dataset. We chose the dataset from the paper, *Electoral Backlash against Climate Policy: A natural Experiment on Retrospective Voting and Local Resistance to Public Policy*, replicated the traditional two-stage least squares regression, and applied our two-stage and one-stage methods.

Here, we gave a brief introduction of this paper and the dataset. This paper investigates whether living closed to a wind energy project leads citizens or residents to vote against an incumbent government due to its climate policy. The dataset consists of the election, census and wind energy project data of 708 valid precincts in Ontario. Each row represents a valid precinct. For each precinct, it includes the average wind power (log) in a precinct as the instrument z, whether there is a proposed wind turbine within 3 km of the precinct in 2011 as the treatment t, and the change in the Liberal Party vote share in that precinct between the 2007 and 2011 elections as outcomes y, and features about geographical information as other covariates x.

Due to the fact that the treatment t is a binary variable, neither the linear regression nor the general additive model is applicable any more. Here, we uses the logistic regression for the binary variable instead in the first stage. Since there is no closed-form solution for the coefficients of the logistic regression, we redo the programming using the general loss version of our methods instead of the square loss.

The actual data analysis procedures are as follows. First, we checked the relevance assumption and the exclusion restriction of the instrument z in this dataset. Since the input features containing the instrument z are significant in the first stage model, the relevance assumption is satisfied and the instrument z has a strong first stage. As our empirical

validity check is passed, the exclusion restriction is also satisfied and the instrument z is valid. Second, we built four different models using different input features and used RMSE as the metric to compare the prediction performance. The results are shown in the table below.

Table 9.1: Comparison of Four Different Models in the Testset (from the simplest to the most complicated)

Models (Input Features)	**Causal Effect**	**Neyman Variance**
Model 1 Linear Regression (with only linear features)	-0.5445	1.3929×10^{-2}
Model 2 Logistic Regression (with only linear features)	-0.5615	1.3819×10^{-2}
Model 3 Logistic Regression (adding non-linear features of the covariates x)	-0.6209	1.3754×10^{-2}
Model 4 Logistic Regression (adding interaction terms between the instrument z and the covariates x, the predicted values of the treatment $\hat{t}$ and the covariates x)	-0.6493	1.4657×10^{-2}

Note that we have rescaled all the input variables in the data pre-processing procedures. Therefore, the results shown in the table above are also standardized.

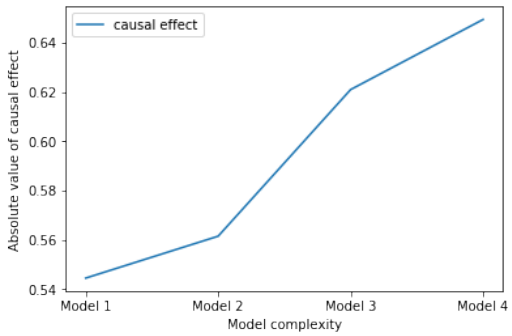

Figure 9.1: Causal effect size v.s. models with different model complexity

According to the visualization above, as the model becomes more complicated, the causal effect size increases. It implies that the traditional two-stage method tends to underestimate the true causal effect with a simple linear framework, while our methods can provide a more accurate estimation.

Chapter 10

Conclusion

The traditional two stage instrumental variable (IV) model, also called the two stage least square regression (2SLS) is design under the linear framework. However, this framework can lead to inaccurate predictions and wrong judgments of the instrument validity due to the construction limitation. In order to break the limitation, this paper general the traditional two stage IV model to the non-linear framework which can provide more choices of model constructions. The usage of the traditional two stage IV model require two critical assumptions, the relevance assumption and the exclusion restriction, which are also defined under the linear framework and use the correlation as the measurement. Due to the fact that the correlation can only measure the linear agreement, it can be problematic under the non-linear framework. Therefore, to be consistent with the general two stage IV model, this paper introduced the corresponding general versions of two critical assumptions. The new versions of assumptions use the prediction loss to measure the relationship between variables, which can detect more complicated dependence. As the new version of the relevance assumption can be assessed directly while the new version of the exclusion restriction cannot, this paper further introduced an empirical validity check for the new version of the exclusion restriction. This empirical validity check using the remainder as an approximation to the error term in the second stage can partially assess the instrument under the non-linear framework. In order to further reduce the computational error, this paper also introduced a new two-stage method and a one-stage method which incorporate the empirical validity check as a constraint into the optimization problems.

In Chapter 6, the limitations of our two-stage method are illustrated and proved. Given some specific model constructions, our empirical validity check can be never active (always satisfied) or impossible to be improved. In these cases, our methods can neither be used to check whether the instrument is valid nor help find better coefficients of the IV models.

It shows that our methods cannot always outperform the traditional two stage IV method. For some specific model settings, the traditional one is still the best possible method.

However, in other cases, our two-stage and one-stage methods do show stronger prediction power than the traditional two-stage IV method. Since our general methods can provide more choices of model constructions, they can better capture the more complicated dependence between variables, which leads to more accurate prediction of the outcomes. As it is shown in Chapter 8, more accurate prediction of the outcomes also leads to more accurate estimation of the causal effect, which is critical in the causal inference analysis.

Note that although our one stage method is more flexible than our two stage method, more flexibility also brings less stability. The flexible solution set in the first stage provided by our one stage method makes it more likely to reach a local optimum. Therefore, in practice, our two stage method is more recommended to ensure stability.

When introducing our two-stage and one-stage methods in Chapter 4, we only considered the square loss function, due to the fact that the least square solution has the closed-form expression, which make the mathematical deduction more feasible. However, the non-linear framework we provided can also be applied on other loss functions. For example, in Chapter 8, we considered the maximum likelihood loss function when using the logistic regression for the binary treatment. As for the future work, we would further generalize our two-stage and one-stage methods to be applied on other loss functions, which would give us even more flexibility in modelling.

Bibliography

[GGHS13] Thomas A. Glass, Steven N. Goodman, Miguel A. Hernn, and Jonathan M. Samet. Causal inference in public health. *Annual Review of Public Health*, 34(1):61–75, 2013. PMID: 23297653.

[KKMS16] Hyunseung Kang, Benno Kreuels, Jrgen May, and Dylan S. Small. Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *Ann. Appl. Stat.*, 10(1):335–364, 03 2016.

[RIHZ00] Donald B. Rubin, Guido W. Imbens, Keisuke Hirano, and Xiao-Hua Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, 03 2000.

[Sto16] Leah C. Stokes. Electoral backlash against climate policy: A natural experiment on retrospective voting and local resistance to public policy. *American Journal of Political Science*, 60(4):958–974, 2016.

[SW11] James Stock and Mark Watson. *Introduction to Econometrics*. Addison Wesley Longman, 3rd edition, 2011.

[SY05] James Stock and Motohiro Yogo. *Testing for Weak Instruments in Linear IV Regression*, pages 80–108. Cambridge University Press, New York, 2005.