Relaxing Instrument Exogeneity with Common Confounders

CHRISTIAN TIEN [*]

August 27, 2023

Abstract

Instruments can be used to identify causal effects in the presence of unobserved confounding, under the famous relevance and exogeneity (unconfoundedness and exclusion) assumptions. As exogeneity is difficult to justify and to some degree untestable, it often invites criticism in applications. Hoping to alleviate this problem, we propose a novel identification approach, which relaxes traditional IV exogeneity to exogeneity conditional on some unobserved common confounders. We assume there exist some relevant proxies for the unobserved common confounders. Unlike typical proxies, our proxies can have a direct effect on the endogenous regressor and the outcome. We provide point identification results with a linearly separable outcome model in the disturbance, and alternatively with strict monotonicity in the first stage. General doubly robust and Neyman orthogonal moments are derived consecutively to enable the straightforward $\sqrt{n}$-estimation of low-dimensional parameters despite the high-dimensionality of nuisances, themselves non-uniquely defined by Fredholm integral equations. Using this novel method with NLS97 data, we separate ability bias from general selection bias in the economic returns to education problem.

---

[*]ct493@cam.ac.uk; Faculty of Economics, University of Cambridge

1   Introduction

Unobserved confounding complicates the identification of a causal effect of a regressor of interest on an outcome. Despite the endogeneity of a regressor of interest, instrumental variable (IV) approaches can identify their causal effect if the famous relevance and exogeneity (unconfoundedness and exclusion) assumptions hold for the instruments. These assumptions are strong and often invite criticism of IV estimates in practice. We propose a novel approach to relax exogeneity, in favour of exogeneity conditional on an unobserved common confounder, for which some relevant variables are observed.

Relaxing exogeneity is only possible when it is replaced by other strong assumptions. One way to identify causal effects without instrument exogeneity is from residual distributions, not variation in the explanatory variables [Heckman, 1979, Millimet and Tchernis, 2013]. Very specific forms of heteroskedasticity across the first stage and outcome model can also be used to establish identification without an exclusion restriction [Klein and Vella, 2010, Lewbel, 2012]. Others have suggested to use irrelevant variation in instruments to test for the exogeneity of the relevant variation in the instruments [D'Haultfœuille et al., 2021]. A similar idea is followed when integrated conditional moments use nonlinear mean-dependence of endogenous variables on instruments, such that the instruments may violate the exclusion restrictions in pre-specified parametric ways. Despite recent advances in estimation with integrated conditional moments [Tsyawo, 2021], the strong identifying assumptions render all these approaches difficult to justify in applications. Our solution differs significantly from these approaches as it only uses a relaxed exogeneity assumption and variation in explanatory variables to identify the causal effect of interest.

From the perspective of IV, we allow for some endogeneity in the instruments. That endogeneity originates from unobserved common confounders. We call these unobserved confounders *common*, because there are some observed variables, which are relevant for them. These observed variables are called proxies. In other words, we assume there are some unobserved variables that explain all correlation (association) between the instruments and these proxies. This assumption is testable. Then, we need to argue for the exogeneity of the instruments, but only conditional on the unobserved common confounders (and observable variables). In general, this is a strong relaxation of instrument exogeneity conditional on observed variables only.

Another way to understand our proposal is as a solution to measurement error in observed confounders for IV. Residual bias is a well-known problem when confounders are measured with error. Proximal learning [Cui et al., 2020] is a solution to the problem, where observed variables measure all unobserved confounders with some error. In proximal learning, the

proxies for the unobserved confounder may either be causes of the treatment or outcome variable. These proxies must be sufficiently relevant (e.g. complete) for all unobserved confounders. Separately developed from the proximal learning approach, a control function solution exists with identical conditional independence assumptions and mismeasured confounders [Nagasawa, 2018]. Our solution is different, as we do not assume the existence of measurements for all confounders. Instead, we use instruments and assume that measurements exist for all confounders, conditional on which the instruments would be exogenous. In this sense, our solution can be understood as IV with mismeasured confounders.

As it is standard in the control function literature, our approach will identify average causal (structural) quantities of interest. Unless the outcome model is fully linearly separable in the treatment and disturbance [Newey et al., 1999], where in our case the disturbance includes the effect of the common confounder, we identify those average causal (structural) quantities of interest that integrate out the unobservables without dependence on the treatment using a control function [Imbens and Newey, 2009]. Our identification approach is most similar to recent advances in nonlinear panels [Liu et al., 2021]. In panel data, unobserved fixed effects are common to the same variables across time, in a similar way as the unobserved common confounders are common to the instruments and proxies in our setup. In Liu et al. [2021], identification stems from a parametric dimension reduction of the effect of observed variables on the outcome, and an index sufficiency assumption that renders the observed variables independent from the fixed effects conditional on an index of the observed variables. In our approach, identification stems from the existence of more instruments than treatments, and an index sufficiency assumption that renders the instruments independent from the unobserved common confounders conditional on an index of the instruments. Just like Blundell and Powell [2004], Liu et al. [2021] do not explain how to derive this crucial index function. One of our main contributions is the derivation of the index function, which arises naturally in the common confounding setup.

A motivating example for our proposal is the returns to college education identification problem. It features various biases, ability and selection, and we motivate pre-college test scores as instruments exogenous to selection, while clearly endogenous to ability. proxies are pre-college risky behaviour dummies, which appear to correlate negatively with ability. With NLS97 data, we show that selection bias is the much more economically relevant bias compared to ability bias in this problem.

2 Setup

The treatment (action) $A \in \mathcal{A}$ is discrete or continuous with base measure μ_A of $\mathcal{A} \subseteq \mathbb{R}^{d_A}$. $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the one-dimensional outcome variable. Other important variables are the instruments $Z \in \mathcal{Z} \in \mathbb{R}^{d_Z}$, the proxies $W \in \mathcal{W} \subseteq \mathbb{R}^{d_W}$, and the common confounders $U \in \mathcal{U} \subseteq \mathbb{R}^{d_W}$.

Assumption 2.1 (Common Confounding IV Model). *1. SUTVA: $Y = Y(A, Z)$.*

2. Instruments

 (a) Exogeneity: $Y(a, z) = Y(a) \perp\!\!\!\perp Z \mid U$.

 (b) Index sufficiency: For some $\tau \in \mathcal{L}_2(Z)$, where $T := \tau(Z)$, $U \perp\!\!\!\perp Z \mid T$.

 (c) Relevance (completeness): For any $g(A, T) \in \mathcal{L}_2(A, T)$,

$$\mathbb{E}\left[g(A, T) | Z\right] = 0 \text{ only when } g(A, T) = 0. \tag{2.1}$$

3. Proxies

 (a) Exogeneity: $W \perp\!\!\!\perp Z \mid U$.

 (b) Relevance (completeness): For any $g(U) \in \mathcal{L}_2(U)$,

$$\mathbb{E}\left[g(U) | W\right] = 0 \text{ only when } g(U) = 0. \tag{2.2}$$

Assumption 2.1.1 is the standard stable unit treatment value assumption (SUTVA), which implies no interference across units. In assumption 2.1.2a we capture the key relaxation of this model compared to standard IV. It states that the instruments are excluded and unconfounded, yet this exclusion and unconfoundedness may be conditional on an unobserved (vector-valued) random variable U. This is a significant relaxation of the standard exclusion restriction and unconfoundedness assumption, which is possible only with assumptions 2.1.2b, 2.1.2c, and 2.1.3. In assumption 2.1.2b, we introduce a control function $\tau \in \mathcal{T} \subseteq \mathcal{L}_2(Z)$ and a control variable $T := \tau(Z)$ such that $T \in \mathcal{Z}^\tau \subseteq \mathcal{Z}$. Conditional on the control variable T, the instruments Z are independent from the common confounders U. A simple example of such a function is the conditional density $f(U|Z)$. This assumption describing the existence the control function τ is often called index sufficiency, where T is a (multiple) index of Z. In assumption 2.1.2c, we require that conditional on the control variable T, the instruments Z are complete for treatment A. This is a standard completeness condition. It simply means that keeping the variation of Z described by T fixed, the

3

instruments must remain sufficiently relevant for A. In slightly different words, after conditioning on T, enough variation must be left in the instruments Z to infer the effect of treatment A on outcome Y. Intuitively, we are orthogonalising Z with respect to some of its own variation, the variation in T. As in standard IV with observed confounders, this relevance requirement is typically testable. Assumption 2.1.3a states that the proxies W are independent from instruments Z conditional on the common confounders U. The proxies W must also be complete for the unobserved common confounders U, as stated in 2.1.3b. Again, completeness means the proxies W are sufficiently relevant for U.

A different way to understand these assumptions is that the unobserved variable U, which explains all association (correlation) between the proxies W and instruments Z, renders the instruments exogenous when observed. In this sense, W can be (possibly quite poor) proxies for what we consider the unobserved common confounder U, as long as they are sufficiently relevant. Conditional on W, the instruments Z are still endogenous. Common confounders U are never observed, and W could be quite poor proxies for it. Yet, we prove that conditioning on a control function T, which renders the instruments Z and proxies W independent, exogeneity of instruments Z is restored.

3   Learning the Confounding Structure

In this section, we describe the main idea of the paper. Using only observable information, we find a control variable T, conditional on which the instruments Z are independent from the unobserved common confounders U. We then explain what may be considered the optimal control variable T.

3.1   Learning a Control Function

The control function $\tau \in \mathcal{L}_2(Z)$, described in lemma 1, generates the control variable T. This control variable renders the instruments Z independent from the unobserved common confounders U. Logically, if the instruments Z and the proxies W are independent conditional on U, it follows that any such control variable T also renders Z and W independent conditional on T.

Lemma 1. *Assume $W \perp\!\!\!\perp Z \mid U$ (2.1.3a). Take any $\tau \in \mathcal{L}_2(Z)$, where $T := \tau(Z)$, such that $U \perp\!\!\!\perp Z \mid T$. Then, also $W \perp\!\!\!\perp Z \mid T$.*

One possible such control variable is $T = Z$, yet this would leave no remaining variation in Z to instrument for A conditional on T. Also, lemma 1 does not provide a way to identify

any control function τ apart from a function which captures the same information as Z itself. For this purpose, we need lemma 2. In this lemma, we establish that any $T = \tau(Z)$, conditional on which the instruments Z and proxies W are independent, also renders Z conditionally independent from the unobserved common confounders U.

Lemma 2. *Assume $W \perp\!\!\!\perp Z \mid U$ (2.1.3a), and for any $g(U) \in \mathcal{L}_2(U)$, $\mathbb{E}\left[g(U)|W\right] = 0$ only when $g(U) = 0$ (2.1.3b). Take any $\tau \in \mathcal{L}_2(Z)$, where $T := \tau(Z)$, such that $W \perp\!\!\!\perp Z \mid T$. Then, also $U \perp\!\!\!\perp Z \mid T$.*

Unlike lemma 1, the conclusion of lemma 2 is not obvious and requires the completeness of proxies W for the unobserved common confounders U. Again, completeness means that the proxies W must be sufficiently relevant for U. If this were not the case, it would be impossible to keep all variation in Z that is associated with U fixed, using a control variable T derived only using information about the association of Z and W. We can interpret U as all unobserved confounders that associate (correlate) Z and W.

Lemma 2 is an important result, because it allows the identification of a control function that does not capture all variation in instruments Z. For any T that we identify conditional on which instruments Z and proxies W are independent, the instrument exogeneity assumption can be relaxed to exogeneity conditional on all unobservables U that associate (correlate) Z and W (assumption 2.1.2a). The parallels to standard IV are quite clear: The conditional exogeneity assumption 2.1.2a is untestable, yet relaxed compared to standard IV. The relevance requirement of Z for A conditional on T (assumption 2.1.2c) is testable, yet stricter compared to standard IV. The requirement for relevance of Z for A conditional on T implies that only a subset of control functions $\tau \in \mathcal{L}_2(Z)$, which leave enough relevant variation in Z conditional on T, enable model identification under assumption 2.1:

$$\mathcal{T}_{\text{valid}} := \left\{\tau \in \mathcal{L}_2(Z): \ (W \perp\!\!\!\perp Z \mid \tau(Z)) \ \text{and} \ \left(\mathbb{E}\left[g(A, \tau(Z)|Z\right] = 0 \ \text{only when} \ g(A, \tau(Z)) = 0\right)\right\}$$

$$(3.1)$$

As both defining relevance conditions of this set $\mathcal{T}_{\text{valid}}$ are testable, its non-emptiness is testable as well.

3.2 Optimal Control Function

Under assumption 2.1, the optimal control function $\tau^* \in \mathcal{T}_{\text{valid}}$ out of the set of valid control functions captures the minimum feasible information in Z in a sense of minimising the variance of the asymptotically unbiased estimator $\hat{\theta}$ of some causal estimand θ. Figure 1

illustrates schematically how the degree of inconsistency and asymptotic variance of an IV estimator conditional on the control variable T depend on the complexity of τ.

Figure 1: Implied typical estimator properties with control functions τ of varying complexity
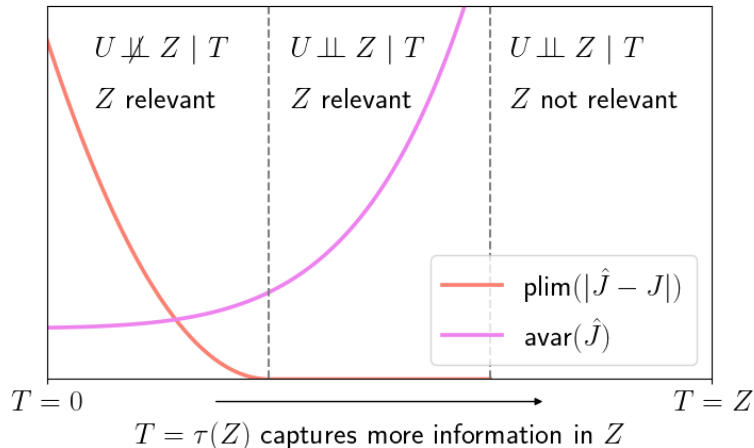
Notes: This figure illustrates some typical properties of an estimator $\hat{\theta}$ of the causal effect of a treatment A on an outcome Y, using instruments Z while conditioning on a control function $T = \tau(Z)$. Moving to the right on the x-axis, the control function captures more information in Z, starting with variation in Z which correlates with the unobserved common confounders U.

In the left rectangle, the control function is too simple to render Z exogenous conditional on T. Hence, the estimator is inconsistent, yet the degree of inconsistency decreases as the complexity of the control function T increases. In the central rectangle, the control function captures enough information for Z to be excluded conditional on T. So, the estimator is consistent. In the right rectangle, the control function is too complex. Conditional on T, the instruments Z are no longer sufficiently relevant for A and the estimator $\hat{\theta}$ asymptotically does not exist. In the first two rectangles, the asymptotic variance of the estimator $\hat{\theta}$ increases with the complexity of the control function, because conditional on T, less information in Z is used to infer the effect of A on Y.

In figure 1, the complexity of τ on the x-axis increases from $\tau(Z) = 0$ on the extreme left to $\tau(Z) = Z$ on the extreme right. Moving further to the right on the x-axis means that the control function captures more information in Z, starting with variation in Z which correlates with the unobserved common confounders U. In the left rectangle, the complexity of τ is low. $T = \tau(Z)$ does not capture all information in Z that correlates with U, so even conditional on T the instruments Z remain endogenous, and the estimator $\hat{\theta}$ is inconsistent. However, as all information in Z is used for inference, the asymptotic variance of the estimator $\hat{\theta}$ will be relatively small. As the complexity of τ increases towards the right in the left quadrant, more information about the elements of Z which correlate with U is captured in T. Increasing the complexity of τ increases the asymptotic variance of $\hat{\theta}$ as less information in Z is used. Importantly, as this information corresponds to variation in Z that correlates

with U, inconsistency is being reduced.

In the central rectangle of figure 1, the complexity of τ is sufficient for Z to be exogenous conditional on T. Hence, the estimator $\hat{\theta}$ is consistent. However, as τ increases in complexity, we use less information in Z to infer the causal estimand θ. Hence, inevitably the asymptotic variance of the estimator $\hat{\theta}$ increases. Consequently, the optimal τ would be that of minimal complexity, such that Z is excluded conditional on T. In practice, we do not know but can only estimate $\mathcal{T}_{\text{valid}}$, so the exact minimum complexity valid τ is unknown and estimated with sampling error. However, even if a τ is chosen with slightly too little complexity, the resulting inconsistency may still be small. The margin of sufficient complexity is at the border of the left and central rectangle. When a τ with slightly too little complexity is chosen, a small degree of inconsistency is incurred, but depending on the sample size possibly outweighed in terms of mean squared error contribution by the associated standard deviation reduction. This is an example of a small in-sample bias-variance tradeoff, while we otherwise focus on identification to enable the construction of consistent estimators.

As the complexity of τ increases, at some point the instruments Z are no longer relevant for treatment A conditional on T. Asymptotically, the estimator $\hat{\theta}$ no longer exists. This is the case in the right rectangle of figure 1. In the extreme, τ is simply an identify function and $T = Z$. No variation in Z remains to infer the effect of A on Y. However, even in less extreme cases where there is some variation left in Z conditional on T, it may simply be insufficient variation to be relevant for A.

3.3 Specification Test

A straightforward way to ensure the sufficient complexity of some τ is to test $W \perp\!\!\!\perp Z \mid T$. An alternative to this is a specification test, similar in spirit to specification testing in overidentified IV models. Consider the two control functions τ_1 and τ_2, such that $T_1 = \tau_1(Z)$ and $T_2 = \tau_2(Z)$. Without loss of generality, let τ_1 be *less* complex than τ_2, i.e. $\mathcal{R}(\tau_1) \subset \mathcal{R}(\tau_2)$ where $\mathcal{R}(a)$ is the range space of function a. The null hypothesis is the conditional exogeneity of Z given T_1,

$$H_0 : Y(a) \perp\!\!\!\perp Z \mid T_1, \text{ with alternative } H_a : Y(a) \not\!\perp\!\!\!\perp Z \mid T_1.$$

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the two causal estimators of estimand θ_0 based on τ_1 and τ_2. Suppose that under both control functions, the instruments Z remain conditionally relevant for treatment A, so that both estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ have some probability limit. We also still assume that conditional on U, the instruments Z are exogenous. The conditional exogeneity of Z given

U is assumed, because here we only test for the sufficient complexity of τ, i.e. $Z \perp\!\!\!\perp U \mid T$.[1]

Under the null hypothesis H_0, both estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ converge to the true causal effect θ_0. However, the asymptotic variance of $\hat{\theta}_2$ with the more complex control function τ_2 will be larger than that of $\hat{\theta}_1$, as $\hat{\theta}_2$ uses less variation in Z than $\hat{\theta}_2$. Under the alternative H_a, the estimators do not have the same probability limit. If τ_2 still captures enough variation T_2 in Z for the instruments Z to be conditionally exogenous, $\hat{\theta}_2$ still converges to θ_0. $\hat{\theta}_1$ on the other hand will no longer converge to θ. Generally, there is no guarantee that T_2 still renders Z conditionally exogenous. In this case, $\hat{\theta}_2$ converges to some value other than θ_0. However, unless the additional variation that we condition on in T_2 compared to T_1 is exogenous due to some particularly poor construction of τ_2, $\hat{\theta}_1$ and $\hat{\theta}_2$ still have different probability limits. A specification test using this logic is generally possible for the sufficient complexity of a control function τ.

4 Point Identification

Without further parametric restrictions on the outcome or first stage model, at most set identification is possible. When the outcome model is linearly separable in the observables and unobservables, we show how to point-identify the model part relating to the observables [Newey and Powell, 2003]. If instead the first stage is monotonous, a control function approach can be used to point-identify average structural functions and thus causal effects with a common support assumption (instead of completeness) [Imbens and Newey, 2009]. We construct a control function for the endogenous variation in A while already keeping the endogenous variation in Z fixed.

4.1 Linearly separable outcome model

An outcome model with linear separability in the treatment and a disturbance is one special case where point identification is possible. With a linearly separated disturbance ε, it is straightforward to represent the exogeneity of instrument Z as mean-independence conditional on common confounders U. Assumption 4.1 fully describes this setting.

Assumption 4.1 (Linearly separable outcome model)**.** *There exists some function $k_0 \in \mathcal{L}_2(A)$ such that*

$$Y = Y(A) = k_0(A) + \varepsilon, \qquad\qquad \mathbb{E}\left[\varepsilon|Z,U\right] = \mathbb{E}\left[\varepsilon|U\right]. \qquad\qquad (4.1)$$

---

[1] To test whether some instruments Z are exogenous conditional on U, we can use a standard specification test for different Z, if θ_0 is overidentified conditional on T.

The conditional moment describes the mean-independence of instruments Z conditional on the unobserved common confounders U.

Theorem 1 (Identification in linearly separable model). *Let assumptions 2.1.(1/2b/2c/3) and 4.1 hold. Any $h \in \mathcal{L}_2(A, T)$ for which $\mathbb{E}[Y|Z] = \mathbb{E}[h(A, T)|Z]$, satisfies $h(A, T) = k_0(A) + \mathbb{E}[\varepsilon|T]$. Consequently, $\theta := \int_\mathcal{A} Y(a)\pi(a)\mathrm{d}a = \int_\mathcal{T} \int_\mathcal{A} h(a, t)\pi(a)\mathrm{d}a f(t)\mathrm{d}t$.*

Theorem 1 establishes point identification of the function k_0 of the effect of the observable treatment A on outcome Y. While we do not make this explicit, k_0 may also be a function of the proxies W or other observed covariates X. The linear separability in combination with the completeness assumption leads to a straightforward identification in the linearly separable model. Unlike in Tien [2022], identification of an average structural function when there are interactions of the observables and unobservable U is much more difficult in this model where the proxies W may also have a direct effect on treatment A.

We could have considered other model specifications or versions of completeness to establish identification [D'Haultfoeuille, 2011]. For now, we leave this exercise for future work.

4.2 First stage monotonicity

If the outcome model is not linearly separable in treatment and disturbance, monotonicity in the first stage reduced form is an alternative assumption to identify average causal (structural) effects [Imbens and Newey, 2009]. If the common confounders U were observed, there would be a simple control function for the endogenous variation in A due to monotonicity. Assumption 4.2 describes the necessary first stage reduced form monotonicity.

Assumption 4.2 (Monotonicity).

$$A = h(Z, \eta) \tag{4.2}$$

1. $h(Z, \eta)$ *is strictly monotonic in η with probability 1.*

2. η *is a continuously distributed scalar with a strictly increasing conditional CDF $F_{\eta|U}$ on the conditional support of η.*

3. $Z \perp\!\!\!\perp \eta \mid U$.

Assumption 4.2.1 describes the strict montonicity of A in the disturbance η. This disturbance η is scalar and continuously distributed conditional on the unobservable U, with a strictly increasing conditional CDF according to assumption 4.2.2. Jointly, these two assumptions ensure that for any given Z, any A is associated with a unique η. The unobserved

confounders U may affect A, but only through their effect on η. This restriction keeps the model monotonous in the unobservables to ensure point identification. Finally, assumption 4.2.3 requires full independence of instruments Z and η conditional on the common confounders U.

The above setup does not immediately help with identification, because the common confounders U are always unobserved. In lemma 3, we establish a few useful facts about the conditional distribution of the scalar disturbance η given T. Notably, this conditional distribution $F_{\eta|T}$ is also strictly increasing on the conditional support of η, and unsurprisingly the instruments Z are independent from η conditional on T.

Lemma 3.

$$F_{\eta|T} := \int_{\mathcal{U}} F_{A|Z,U}(A, Z, u) f_{U|T}(u, T) \, \mathrm{d}\mu_U(u)$$

is a strictly increasing CDF on the conditional support of η, and $Z \perp\!\!\!\perp \eta \mid T$.

The above lemma 3 implies that $F_{\eta|T}(\eta)$ is a one-to-one function of η conditional on T, just like $F_{\eta|U}(\eta)$ conditional on U. This fact is useful, because it is no longer necessary to condition on the unobservable U to identify the endogenous variation in A, which is η, exactly. Instead, if we can identify $F_{\eta|T}(\eta)$, η is held fixed as long as $(F_{\eta|T}(\eta), T)$ is held fixed. The remaining difficulty is to identify $F_{\eta|T}(\eta)$. In this regard, theorem 2 states that $F_{\eta|T}(\eta)$ is equal to the conditional CDF of A given Z. This conditional CDF is defined as V_T in equation 4.3.

Theorem 2. *Let*

$$V_T := F_{A|Z;T}(A, Z). \tag{4.3}$$

Under assumption 4.2, $V_T = F_{\eta|T}(\eta)$, and

$$A \perp\!\!\!\perp Y(a) \mid (V_T, T), \text{ for all } a \in \mathcal{A}. \tag{4.4}$$

Theorem 2 states that despite the unobservable common confounder U, there exist the observable control functions V_T and T conditional on which we retrieve unconfoundedness. Specifically, we retrieve unconfoundedness because all variation in treatment A stems from instruments Z once we condition on V_T and T. Fortunately, conditional on T, the instruments Z are fully exogenous. So far, our arguments have only been with respect to exogeneity, not yet relevance.

To describe relevance, we use a common support assumption 4.3, with focus on a causal effect of interest

$$\theta_0 := \int_{\mathcal{A}} Y(a)\pi(a) \,\mathrm{d}\mu_A(a).$$

This common support assumption requires the sufficient relevance of instruments Z for treatment A, and sufficient variation in Z, both conditional on T. In slightly different words, after holding all variation in Z associated with U fixed through T, the variation in Z must still be sufficiently rich and relevant for A.

Assumption 4.3 (Common Support)**.** *For all $a \in \mathcal{A}$, where the contrast function is non-zero ($\pi(a) \neq 0$), the support of (V_T, T) equals the support of (V_T, T) conditional on A.*

With the common support assusmption 4.3, average causal quantities θ_0 in our model are identified under monotonicity (assumption 4.2). In theorem 3, we explicitly replace the completeness assumption in assumption 2.1.2c by the common support assumption 4.3, which is the correct relevance requirement with a control function.

Theorem 3 (Average causal quantity identification)**.** *Suppose assumption 2.1.(1/2a/2b/3) [relaxed IV model], 4.2 [monotonicity], and 4.3 [common support] hold. Then, any $\theta_0 := \int_{\mathcal{A}} Y(a)\pi(a) \,\mathrm{d}\mu_A(a)$ is identified as*

$$\theta_0 = \int_{\mathcal{V}_T, \mathcal{Z}^\tau} \int_{\mathcal{A}} \mathbb{E}\left[Y | A = a, (V_T, T) = (v_T, t)\right] \pi(a) \,\mathrm{d}\mu_A(a) \,\mathrm{d}F_{V_T, T}(v_T, t).$$

We simply integrate out the control functions (V_T, T) without dependence on treatment A to obtain the causal quantity of interest θ_0. Often, θ_0 will be some form of average treatment effect.

Other functions of interest than the above (weighted) averages of potential outcomes, e.g. quantile structural functions, are also identified as a consequence of theorem 2, but require corresponding common support assumption which will differ from assumption 4.3.

5   Linear Model

In this section, we explain identification in the common confounding model in linear terms. Apart from the illustrative purpose of linear models, their tractability and ease of use make them attractive. For the common confounding IV approach, the linear model provides useful intuition regarding the relevance and exogeneity assumptions.

First, we describe the model assumptions in linear form. The outcome variable $Y \in \mathbb{R}$ is one-dimensional. For ease of notation, we let $A \in \mathbb{R}$ be one-dimensional too. All other

variables X are of some general dimensions d_X, i.e. $Z \in \mathbb{R}^{d_Z}$, $W \in \mathbb{R}^{d_W}$, and $U \in \mathbb{R}^{d_U}$. As previously, instruments are called Z, proxies W, and the unobserved common confounders U.

$$Y = A\beta + U\gamma_Y + W\upsilon_Y + \varepsilon_Y, \qquad \mathbb{E}\left[\varepsilon_Y^\mathsf{T}(Z, U, W)\right] = \mathbf{0} \qquad (5.1)$$

$$A = Z\zeta + U\gamma_A + W\upsilon_A + \varepsilon_A, \qquad \mathbb{E}\left[\varepsilon_A^\mathsf{T}(Z, U, W)\right] = \mathbf{0} \qquad (5.2)$$

Equation 5.1 expresses Y as a linear function of A, U, W and a disturbance ε_Y. The disturbances ε_Y is uncorrelated from the instruments Z. With this moment equation, the model parameters would be identified under a conditional relevance requirement of Z for A if U were observed. The parameter vector of interest in this model is β, the effect of treatment A on Y. In equation 5.2, A is a linearly projected on Z, U, W, and a disturbance ε_A. The $(d_Z \times d_A)$-dimensional parameter matrix ζ describes the marginal linear effect of Z on A. Equation 5.2 for A is the model's first stage. The conditional relevance requirement of Z for A would simply be $\operatorname{rank}(\zeta) = d_A$, if U were observed. With observable U, the model would be sufficiently described at this point to point identify β. As the common confounders U are never observed, the model requires further assumptions.

$$Z = U\gamma_Z + \varepsilon_Z, \qquad \mathbb{E}\left[\varepsilon_Z^\mathsf{T}(U, W)\right] = \mathbf{0} \qquad (5.3)$$

$$W = U\gamma_W + \varepsilon_W, \qquad \mathbb{E}\left[\varepsilon_W^\mathsf{T}(U, Z)\right] = \mathbf{0} \qquad (5.4)$$

Equations 5.3 and 5.4 imply that all correlation between Z and W stems from the unobserved common confounders U. There is no direct effect from either on the other. If there were, we could model this by increasing the dimension of U by the corresponding element of Z or W until Z and W are uncorrelated conditional on U. Consequently, the covariance matrix for Z and W is

$$\Sigma_{ZW} = \gamma_Z^\mathsf{T} \Sigma_U \gamma_W, \qquad \operatorname{rank}(\Sigma_{ZW}) = d_U.$$

5.1 First projection

To deal with the unobservedness of U, we use two additional projections compared to standard IV. First, we project W on Z as

$$W = Z\Sigma_Z^{-1}\Sigma_{ZW} + \tilde{\varepsilon}_W = \gamma_Z^\mathsf{T}\Sigma_U\gamma_W + \tilde{\varepsilon}_W, \qquad \mathbb{E}\left[\tilde{\varepsilon}_W Z\right] = \mathbf{0}.$$

Then, choose any $C_Z \in \mathbb{R}^{d_Z \times d_U}$ and $C_W \in \mathbb{R}^{d_W \times d_U}$ such that $C_Z C_W^\intercal = \Sigma_{ZW}$ to create the control variable

$$T := Z\Sigma_Z^{-1} C_Z, \qquad\qquad \Sigma_{ZT} = C_Z, \qquad\qquad \Sigma_T = C_Z^\intercal \Sigma_Z^{-1} C_Z.$$

The control variable T is a lower-dimensional representation of Z, which captures all correlation between Z and W.

5.1.1 Why can this projection help deconfound Z?

All endogeneity in instruments Z stems from their correlation with U. Hence, if the linear projection of U on Z, $Z\Sigma_Z^{-1}\Sigma_{ZU}$, is unchanged for some linear combinations of Z, those linear combinations of Z become excluded instruments. Assumption 5.1 describes the necessary relevance condition for proxies W with respect to U.

Assumption 5.1 (Relevance conditios for γ_W)**.**

$$d_W \geq d_U \tag{5.5}$$

As long as the proxies W are relevant for U in form of assumption 5.1, the endogeneity inducing unobserved covariance Σ_{ZU} is proportional to the observable covariance Σ_{ZW}. As the rank of Σ_{ZW} is d_U, keeping the d_U-dimensional control variable $T = Z\Sigma_Z^{-1}C_Z$ fixed restores instrument exogeneity in Z. Of course, d_U is the minimal sufficient dimension to construct an exogeneity restoring control variable T. Larger dimensions of T are similarly admissible choices to restore instrument exogeneity, but might complicate relevance.

5.2 Second projection

In the second projection step, the instruments are orthogonalised with respect to the control variable T and thus with respect to the linear projection of U on Z. The orthogonalised instrument $\tilde{Z}$ is constructed as

$$\tilde{Z} = Z \underbrace{\left(I_{d_Z} - \Sigma_Z^{-1} C_Z \left(C_Z^\intercal \Sigma_Z^{-1} C_Z \right)^{-1} C_Z^\intercal \right)}_{:=M_{ZW}, \text{ note that } C_Z^\intercal M_{ZW}=\mathbf{0}} D_Z,$$

for some $D_Z \in \mathbb{R}^{d_U \times d_{\tilde{Z}}}$ such that $\text{rank}\left(\left(I_{d_Z} - \Sigma_T^{-1}\Sigma_{TZ}\right)D_Z\right) = d_{\tilde{Z}} \geq d_A$. Importantly, the orthogonalised instrument $\tilde{Z}$ is uncorrelated with U and W:

$$\Sigma_{W\tilde{Z}} = \Sigma_{WZ}M_{ZW}D_Z = C_W \underbrace{C_Z^\intercal M_{ZW}}_{=0} D_Z = 0,$$

$$\begin{aligned}\Sigma_{U\tilde{Z}} = \Sigma_{UZ}M_{ZW}D_Z &= \left(\gamma_W\gamma_W^\intercal\right)^{-1}\gamma_W\Sigma_{WZ}M_{ZW}D_Z \\ &= \left(\gamma_W\gamma_W^\intercal\right)^{-1}\gamma_W C_W \underbrace{C_Z^\intercal M_{ZW}}_{=0} D_Z = 0.\end{aligned}$$

Finally, using equations 5.1 and 5.2 it is easy to show that β can be written in a familiar form:

$$\beta = \mathbb{E}\left[A^\intercal P_{\tilde{Z}}A\right]^{-1}\mathbb{E}\left[A^\intercal P_{\tilde{Z}}Y\right], \tag{5.6}$$

$$P_{\tilde{Z}} := \tilde{Z}\left(\tilde{Z}^\intercal\tilde{Z}\right)^{-1}\tilde{Z}^\intercal = ZM_{ZW}D_Z\left(D_Z^\intercal\Sigma_Z M_{ZW}D_Z\right)^{-1}D_Z^\intercal M_{ZW}^\intercal Z^\intercal. \tag{5.7}$$

5.3 Enhanced instrument relevance condition

With the conditional exogeneity of Z sufficiently discussed, the focus shifts to the conditional relevance of Z for A.

Given that at least d_U dimensions of variation in Z are held fixed, there can be spare variation in Z to identify the linear parameter vector β. A relevant condition for this is $d_Z - d_U \geq d_A$. After keeping the d_U-dimensional variation in $\mathbb{E}[W|Z]$ fixed, the expected predicted values of treatment A given instruments Z, $\mathbb{E}\left[Z\zeta|\mathbb{E}[W|Z]\right]$, must be non-degenerate. The rank condition for any d_A is described in assumption 5.2.

Assumption 5.2 (Rank condition). *For some $C_Z \in \mathbb{R}^{d_Z \times d_U}$ and $C_W \in \mathbb{R}^{d_W \times d_U}$ such that $C_Z C_W^\intercal = \Sigma_{ZW}$, and $D_Z \in \mathbb{R}^{d_U \times d_{\tilde{Z}}}$,*

$$\text{rank}\left(\mathbb{E}\left[A^\intercal P_{\tilde{Z}}A\right]\right) = d_A, \tag{5.8}$$

where

$$P_{\tilde{Z}} := ZM_{ZW}D_Z\left(D_Z^\intercal\Sigma_Z M_{ZW}D_Z\right)^{-1}D_Z^\intercal M_{ZW}^\intercal Z^\intercal,$$

$$M_{ZW} := I_{d_Z} - \Sigma_Z^{-1}C_Z\left(C_Z^\intercal\Sigma_Z^{-1}C_Z\right)^{-1}C_Z^\intercal.$$

d_U dimensions of variation in Z are lost by conditioning on $\mathbb{E}[W|Z]$. The remaining variation in Z after this conditioning step must be sufficiently relevant for A.

6 Semiparametric estimation

With its transparency, the linear model sheds light on the assumptions in IV with common confounders. The intuition for relevance and exogeneity assumptions in the linear model carries on to nonlinear models, specifically the idea of a pre-IV control function. In the linear model, instrument variation is used conditional on the d_U-dimensional control variable T, which is derived from the correlation of Z and W. In nonlinear settings, the control function is a general $\tau \in \mathcal{T} \subseteq \mathcal{L}_2(Z)$, which serves the same purpose: to render the instruments Z conditionally independent from the proxies W and hence from the unobserved common confounders U. Exogeneity of the instruments Z conditional on this control function is the desired consequence.

Generally, we would want to identify control function τ via the integral equation

$$\mathbb{E}\left[g_0(Z)|W\right] = \mathbb{E}\left[\tau(Z)|W\right] \tag{6.1}$$

for some function $g_0 \in \mathcal{G} \subseteq \mathcal{L}_2(Z)$ such that $g_0(Z) \in \mathcal{Z}_0^g$. It is worth considering the behaviour of equation 6.1 for two extreme cases.

First, consider $Z \perp\!\!\!\perp W$, the unconditional independence of instruments and proxies. In the unconditional independence setting, no information in Z must be held fixed to recover exogeneity. Formally, it follows that $\mathbb{E}\left[g_0(Z)|W\right] = \mathbb{E}\left[g_0(Z)\right]$. The function τ does not require its input Z, and can simply be set to $\tau = \mathbb{E}\left[g_0(Z)\right]$.

Secondly, one may note that one extreme solution always solves equation 6.1, that solution being $\tau(Z) = g_0(Z)$. Intuitively, the remaining variation in Z has to be exogenous when the entirety of Z is held fixed, because there simply is no remaining variation in Z.

Any interesting solution will be non-extreme, in the sense that $Z \perp\!\!\!\perp W \mid U$ for unobservable U, and $\tau \neq g_0$. The relevance of that spare information for treatment A is generally testable, as in standard IV conditional on covariates.

Let $\mathcal{T}_{\text{exog}}$ be the set of solutions to 6.1 for a given function $g_0 \in \mathcal{G}$, i.e.

$$\mathcal{T}_{\text{exog}} = \left\{\tau \in \mathcal{T} : \mathbb{E}\left[g_0(Z)|W\right] = \mathbb{E}\left[\tau(Z)|W\right]\right\}. \tag{6.2}$$

As discussed, the $\tau = g_0$ is always in $\mathcal{T}_{\text{exog}}$, but is incompatible with instrument relevance. In a sense, $\mathcal{T}_{\text{exog}}$ is the set of exogeneity restoring control functions. A further relevance assumption will lead to the set of valid control functions $\mathcal{T}_{\text{valid}} \subset \mathcal{T}_{\text{exog}}$. As the required relevance assumptions always depend on the estimand, we will leave this general on purpose, for now.

6.1 Continuous linear functionals in τ

Subsequently, we broadly follow the semiparametric estimation setup in Bennett et al. [2022]. Consider a general semiparametric model where the target estimand θ_0 is given by an unconditional moment as

$$\theta_0 = \mathbb{E}\left[m(O; \tau_0)\right], \ \forall \tau_0 \in \mathcal{T}_{\text{valid}}, \tag{6.3}$$

where m is a known function, and $O = (Y, A, Z, W)$ a vector of all observable variables. Non-uniqueness of valid control functions $\tau_0 \in \mathcal{T}_{\text{valid}}$ is not a problem here, because the target estimand θ_0 remains uniquely identified (as in overidentified IV models).

To facilitate inference, we assume that $\tau \mapsto \mathbb{E}\left[m(O; \tau)\right]$ is a continuous linear functional over $\mathcal{T}$. Continuity and linearity will permit using the Riesz representation theorem [Luenberger, 1997]. Specifically, there exists a Riesz representer α of our functional such that

$$\mathbb{E}\left[m(O; \tau)\right] = \mathbb{E}\left[\alpha(Z)\tau(Z)\right] \ \forall \tau \in \mathcal{T}. \tag{6.4}$$

Then, we can use ideas from automated debiased machine learning to find a quantity that we can use in place of the Riesz representer to debias the plug-in estimator $\mathbb{E}_n\left[m(O; \hat{\tau})\right]$ and retrieve $\sqrt{n}$-convergence along with asymptotic normality [Chernozhukov et al., 2022]. These desired asymptotic properties require functional strong identification 6.1.

Operator language facilitates writing the strong identification definition 6.1, so let $P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}(g) : \mathcal{T} \mapsto \mathcal{L}_2(W)$ be the linear operator given by $[P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}\tau](W) = \mathbb{E}\left[\tau(Z)|W\right]$. Let $P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)} : \mathcal{L}_2(Z) \mapsto \mathcal{T}$ be the adjoint of $P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}$ given by $[P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}q](Z) = \Pi_{\mathcal{T}}\,\mathbb{E}\left[q(W)|Z\right] = \Pi_{\mathcal{T}}\left[q(W)|Z\right]$ for any $q \in \mathcal{L}_2(W)$ and projection operator $\Pi_{\mathcal{T}}[.|Z]$ onto $\mathcal{T}$.

Definition 6.1 (Strong identification of θ_0). θ_0 is strongly identified if $\alpha \in \mathcal{N}^{\perp}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}})$, i.e.

$$\Xi_0 \neq \emptyset, \qquad \text{where} \qquad \Xi_0 := \arg\min_{\xi \in \mathcal{T}}\left(\frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\xi(Z)|W\right]^2\right] - \mathbb{E}\left[m(O; \xi)\right]\right) \tag{6.5}$$

$$= \left\{\xi \in \mathcal{T} : P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}\xi = \alpha\right\}. \tag{6.6}$$

Theorem 4 (Strong identification of θ_0). *Suppose* $\Pi_{\mathcal{T}}\left[q(W)|Z\right] = \Pi_{\mathcal{T}}\left[\mathbb{E}\left[q(W)|U\right]|Z\right]$ *for any* $q \in \mathcal{L}_2(W)$ *and* $\mathcal{T} \subseteq \mathcal{L}_2(Z)$, *and assumption 2.1.3b (completeness of W for U) hold. Also, suppose that instruments Z satisfy relevance conditional on any* $\tau_0 \in \mathcal{T}_{valid} \neq \emptyset$ *such that* $\theta_0 = \mathbb{E}\left[m(O; \tau_0)\right]$, *where* $\tau \mapsto \mathbb{E}\left[m(O; \tau)\right]$ *is a continuous and linear functional over* $\mathcal{T}$.

Then, θ_0 is strongly identified with

$$\alpha \in \mathcal{N}^{\perp}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}}) = \mathcal{N}^{\perp}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)} P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}). \tag{6.7}$$

Definition 6.2 (Debiased representation of θ_0)**.** The debiasing nuisance is

$$q_0(W) = \mathbb{E}\left[\xi_0(Z)|W\right] \ \forall \xi_0 \in \Xi_0. \tag{6.8}$$

The set of valid debiasing nuisances is

$$\mathcal{Q}_0 := \left\{ q \in \mathcal{L}_2(W) : \left[P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)} q\right](Z) = \alpha(Z) \right\}. \tag{6.9}$$

The debiased (Neyman orthogonal) moment ψ with $\theta_0 = \mathbb{E}\left[\psi(O;\tau_0,q_0)\right]$ for any $\tau_0 \in \mathcal{T}_{\text{valid}}$ is

$$\psi(O;\tau,q) := m(O;\tau) + q(W)\left(g_0(Z) - \tau(Z)\right) \tag{6.10}$$

Theorem 5. *Suppose θ_0 is strongly identified. Then, $\theta_0 = \mathbb{E}\left[\psi(O;\tau_0,q_0)\right]$ for any $\tau_0 \in \mathcal{T}_{valid}$ and $q_0 \in \mathcal{Q}_0$. For any $\tau \in \mathcal{T}$, $q \in \mathcal{L}_2(W)$, $\tau_0 \in \mathcal{T}_{valid}$, $q_0 \in \mathcal{Q}_0$,*

$$\mathbb{E}\left[\psi(O;\tau,q)\right] - \theta_0 = -\mathbb{E}\left[(q(W) - q_0(W))(\tau(Z) - \tau_0(Z))\right].$$

This implies error bound

$$\left|\mathbb{E}\left[\psi(O;\tau,q)\right] - \theta_0\right| = \left|\left\langle P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}(\tau - \tau_0), q - q_0\right\rangle\right| = \left|\left\langle \tau - \tau_0, P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}(q - q_0)\right\rangle\right|$$

$$\leq \min\left\{\left\|P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}(\tau - \tau_0)\right\|_2 \|q - q_0\|_2, \ \|\tau - \tau_0\|_2 \left\|P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}(q - q_0)\right\|_2\right\}$$

$$\tag{6.11}$$

Double robustness is satisfied as $\theta_0 = \mathbb{E}\left[\psi(O;\tau_0,q)\right] = \mathbb{E}\left[\psi(O;\tau,q_0)\right]$ for any $\tau_0 \in \mathcal{T}_{valid}$, $\tau \in \mathcal{T}$, $q_0 \in \mathcal{Q}_0$, $q \in \mathcal{L}_2(W)$.
Neyman orthogonality is satisfied as $\tau_0 \in \mathcal{T}_{valid}$ and $q_0 \in \mathcal{Q}_0$ if and only if

$$0 = \left.\frac{\partial}{\partial t}\mathbb{E}\left[\psi(O;\tau_0 + t\tau, q_0)\right]\right|_{t=0} = \left.\frac{\partial}{\partial t}\mathbb{E}\left[\psi(O;\tau, q_0 + tq)\right]\right|_{t=0} \ \forall \tau \in \mathcal{T}, q \in \mathcal{L}_2(W).$$

If we let $\mathcal{T} = \mathcal{L}_2(Z)$ as we usually would to allow for general τ, the condition $P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)} P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}} \xi_0 = \alpha$ for any $\xi \in \Xi_0$ simplifies to $\mathbb{E}\left[\mathbb{E}\left[\xi_0(Z)|W\right]|Z\right] = \alpha(Z)$. The Riesz representer is therefore

closely related to the debiasing nuisance $\xi_0 \in \Xi_0$. More precisely, any $\xi_0 \in \Xi_0$ is equal to the Riesz representer after two projection steps. This double projection property reflects our intuition from the linear model in section 5, where similarly two projection steps were required to restore instrument exogeneity: First proxies W were projected onto instruments Z to retrieve control T, then Z was orthogonalised with respect to T.

The simpler estimation problem as a result of 4 has the additional advantage that estimation error in τ, identified by a possibly ill-posed inverse problem, no longer determines asymptotic behaviour of an estimator dependent on it. Instead, asymptotic behaviour is determined by the estimation error of the projection of τ on W. The projection eliminates the problem from ill-posedness, which can imply a discontinuous dependence of solutions to equation 6.1 on the data distribution [Carrasco et al., 2007].

6.2 Consecutive debiasing of more general functionals of the treatment

Most functionals of interest will not be linear functionals of the debiasing function τ. However, semiparametric estimation of such functionals often remains possible. For example, let the target estimand θ_0 be given by

$$\theta_0 = \mathbb{E}\left[m(O; k_0(\tau_0, g_0))\right], \ \ \forall k_0(\tau_0, g_0) \in \mathcal{K}_0, \ \tau_0(g_0) \in \mathcal{T}_{\text{valid}}. \tag{6.12}$$

Let the set of functions $\mathcal{K}_0(\tau, g)$ be defined as the set of all functions $k_0(\tau, g) : \mathcal{A} \times \mathcal{Z}^\tau \mapsto \mathbb{R}$ that satisfy the first order (Fredholm) integral equation

$$\mathbb{E}\left[k_0(A; \tau, g)|Z\right] = g(Z) - \tau(Z) \text{ for any given } \tau \in \mathcal{T} \text{ and } g \in \mathcal{G}, \tag{6.13}$$

where the dependence of k_0 on is τ and g is made explicit. Then, let $\mathcal{K}_0 := \mathcal{K}_0(\tau_0, g_0)$ for any $\tau_0 \in \mathcal{T}_{\text{valid}}$ and a given g_0 be the set of functions k_0 that identify θ_0. Remember that $\mathcal{T}_{\text{valid}} \subset \mathcal{T}_{\text{exog}}$, where the set $\mathcal{T}_{\text{exog}}$ itself is defined all solutions to the first order integral equation 6.2. In many cases, instead of being a known function, g_0 will also need to be estimated, e.g. if $g_0(Z) = \mathbb{E}\left[g_Y(Y)|Z\right]$ for some known function $g_Y \in \mathcal{L}_2(Y)$.

6.2.1 Common nuisance of type 6.13 for causal questions

With averages over outcomes being the most common estimand, it is worth understanding 6.13 better. Indeed, consider $g_0(Z) = \mathbb{E}\left[Y|Z\right]$ and note that $\mathbb{E}\left[Y|Z\right] = \mathbb{E}\left[\mathbb{E}^{C(U|\tau_0(Z))}\left[Y|A, \tau_0(Z)\right]|Z\right]$, where $\mathbb{E}^{C(U|\tau_0(Z))}\left[Y|A, \tau_0(Z)\right] := \int_{\mathcal{U}} \mathbb{E}\left[Y|A, \tau_0(Z), U = u\right] d(F_{U|\tau_0(Z)}(u|\tau_0(Z)))$, due to the

conditional completeness of Z for $(A, \tau(Z))$ and $\tau_0(Z) = \mathbb{E}\left[Y | \tau_0(Z)\right]$, such that

$$g_0(Z) - \tau_0(Z) = \mathbb{E}\left[\mathbb{E}^{C(U | \tau_0(Z))}\left[Y | A, \tau_0(Z)\right] - \mathbb{E}\left[Y | \tau_0(Z)\right] | Z\right]. \tag{6.14}$$

Now, we perform a hypothetical exercise. Decompose variation in Z into two independent components $\tau_0(Z)$ and $\tilde{\tau}_0(Z)$, i.e. $\tau_0(Z) \perp\!\!\!\perp \tilde{\tau}_0(Z)$. Neither must $\tilde{\tau}$ be estimated in practice, nor is it is not necessary to be able to reconstruct Z from $\tau_0(Z)$ and $\tilde{\tau}_0(Z)$ exactly. In this hypothetical exercise, we should ensure that no spare variation in Z is predictive of $\mathbb{E}^{C(U | \tau_0(Z))}\left[Y | A, \tau_0(Z)\right]$, i.e.

$$\mathbb{E}\left[\mathbb{E}^{C(U | \tau_0(Z))}\left[Y | A, \tau_0(Z)\right] | Z\right] = \mathbb{E}\left[\mathbb{E}^{C(U | \tau_0(Z))}\left[Y | A, \tau_0(Z)\right] | \tau_0(Z), \tilde{\tau}_0(Z)\right]. \tag{6.15}$$

In other words, $\tilde{\tau}_0(Z)$ should capture all variation in Z that predicts $\mathbb{E}^{C(U | \tau_0(Z))}\left[Y | A, \tau_0(Z)\right]$, while being independent of $\tau_0(Z)$.

Note that any $k_0(\tau_0, g_0) \in \mathcal{K}_0(\tau_0, g_0)$ that satisfies $\mathbb{E}\left[k_0(A; \tau_0, g_0) | Z\right] = g_0(Z) - \tau_0(Z)$ will also satisfy

$$\mathbb{E}\left[k_0(A) | \tilde{\tau}_0(Z)\right] = \mathbb{E}\left[g_0(Z) - \tau_0(Z) | \tilde{\tau}_0(Z)\right]. \tag{6.16}$$

Using the independence of $\tau_0(Z)$ and $\tilde{\tau}_0(Z)$, the right-hand side of the above equation reduces to a form of counterfactual mean deviation

$$\mathbb{E}^{C(U, \tau_0(Z))}\left[Y - c \mid A\right] := \int_{\mathcal{Z}^\tau} \left(\mathbb{E}^{C(U | \tau_0(Z))}\left[Y | A, T = t\right] - \mathbb{E}\left[Y | T = t\right]\right) d(F_{\tau_0(Z)}(t)) \tag{6.17}$$

$$= \int_{\mathcal{Z}^\tau \times \mathcal{U}} \left(\mathbb{E}\left[Y | A, T = t, U = u\right] - \mathbb{E}\left[Y | T = t, U = u\right]\right) d(F_{\tau_0(Z), U}(t, u)), \tag{6.18}$$

where $c \in \mathbb{R}$ is some constant. Indeed, this counterfactual mean deviation integrates out all endogenous variation independently from treatment A. This exact desired behaviour allows the identification of typical causal effects, like average treatment effects, making $\mathbb{E}^{C(U, \tau_0(Z))}\left[Y - c \mid A\right]$ an interesting nuisance function to identify causal effects. Given that any k_0 that satisfies 6.13 also satisfies 6.16, the not necessarily unique k_0 to satisfy 6.13 is $k_0(A) = \mathbb{E}^{C(U, \tau_0(Z))}\left[Y - c \mid A\right]$. Despite identification only up to the constant $c \in \mathbb{R}$, causal effects based on this nuisance function are often unique, e.g. if $m_0(O; k) = k_0(a') - k_0(a)$ for two distinct $a, a' \in \mathcal{A}$.

6.2.2 Consecutive debiasing

Consecutive debiasing will allow us to address each first order bias in successive order. Again, assume that $k \mapsto \mathbb{E}\left[m(O;k)\right]$ for $k \in \mathcal{K}$ is a continuous linear functional over $\mathcal{K}$, such that by the Riesz representation theorem

$$\mathbb{E}\left[m_0(O;k)\right] = \mathbb{E}\left[\alpha_{k,0}(A)k(A)\right] \quad \forall k \in \mathcal{K}. \tag{6.19}$$

Assumption 6.1 (Strong instrument relevance). $\alpha_{k,0} \in \mathcal{N}^{\perp}\left(P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)} P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\right)$, *i.e.*

$$\Xi_{k,0} \neq \emptyset, \qquad where \qquad \Xi_{k,0} := \arg\min_{\xi_k \in \mathcal{K}} \left(\frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\xi_k(A)|Z\right]^2\right] - \mathbb{E}\left[m_0(O;\xi_k)\right]\right) \tag{6.20}$$

$$= \left\{\xi_k \in \mathcal{K} : P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)} P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}} \xi_k = \alpha_{k,0}\right\}. \tag{6.21}$$

The strong instrument relevance assumption 6.1 imposes a stronger regularity condition on the smoothness of $\alpha_{k,0}$ than needed for the simple identification of θ_0, because $\theta_0 = \mathbb{E}\left[q_{k,0}(Z)k_0(A;\tau_0,g_0)\right]$ where $\mathbb{E}\left[q_{k,0}(Z)\right] = \alpha_{k,0}(A)$ only requires $\alpha_{k,0} \in \mathcal{R}\left(P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\right)$ as opposed to $\alpha_{k,0} \in \mathcal{R}\left(P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}} P_{A,\mathcal{K}}^{\mathcal{L}_2(Z))}\right) \subseteq \mathcal{R}\left(P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\right)$. The strong instrument relevance in assumption 6.1 ensures $\sqrt{n}$-estimability of the target functional while allowing for ill-posedness in the inverse problem 6.13 which defines any valid k_0 [Bennett et al., 2022].

Strong instrument relevance as stated in 6.1 requires the remaining variation in Z to be relevant enough with respect to outcome-relevant variation in treatment A after integrating out any of the outcome-relevant variation in A associated with the endogenous variation $\tau(Z)$. Hence, the function space $\mathcal{K}$ contains functions of A, where any outcome-relevant variation associated with the possibly endogenous instrument variation $\tau(Z)$ is explicitly subtracted and integrated out without dependence on treatment A. We can understand the function spaces $\mathcal{K}$ and $\mathcal{T}$ (as well as the $\mathcal{K}_0$ and $\mathcal{T}_{\mathrm{exog}}$) as jointly defined:

$$\mathcal{K} \times \mathcal{T} = \left\{(k,\tau) : \left(\mathbb{E}\left[k(A)|Z\right] = g(Z) - \tau(Z)\right) \wedge \left(\mathbb{E}\left[g(Z)|W\right] = \mathbb{E}\left[\tau(Z)|W\right]\right) \ \forall g \in \mathcal{G}\right\} \tag{6.22}$$

$$\mathcal{K}_0 \times \mathcal{T}_{\mathrm{exog}} = \left\{(k,\tau) : \left(\mathbb{E}\left[k(A)|Z\right] = g_0(Z) - \tau(Z)\right) \wedge \left(\mathbb{E}\left[g_0(Z)|W\right] = \mathbb{E}\left[\tau(Z)|W\right]\right)\right\} \tag{6.23}$$

The joint definition in 6.22 illustrates the close relation between the function spaces $\mathcal{K}$ and $\mathcal{T}$. Assumptions like 6.1 are typically interpreted as smoothness restrictions on the Riesz representer $\alpha_{k,0}$. Here, smoothness means that the nuisance $k \in \mathcal{K}$ may only be a func-

tion of treatment A to the degree that remaining instrument variation after integrating out $\tau(Z)$ remains strongly relevant, hence the assumption's name *strong instrument relevance.* Consequently, strong instrument relevance as stated in 6.1 always implies a restriction on the complexity of $\mathcal{T}$. If $\mathcal{T}$ is too encompassing, $\alpha_{k,0} \in \mathcal{N}^{\perp}\left(P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)} P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\right)$ is less likely to be satisfied. In the extreme, $\mathcal{G} = \mathcal{T}$ implies $\mathbb{E}\left[k(A; \tau, g)|Z\right] = 0$, requiring $\alpha_{k,0} = 0$ under strong instrument relevance 6.1, which means that θ_0 is identifiable only if $k = 0$. Clearly, zero nuisances can only correspond to trivial identification problems. For any non-trivial identification problem, the strong instrument relevance assumption 6.1 always implies some restriction on the complexity of $\mathcal{T}$. A violation of strong instrument relevance 6.1 can be interpreted in two ways: The first interpretation is that remaining instrument variation after integrating out $\tau(Z)$ is not sufficiently relevant for treatment A, the alternative interpretation is that too much variation in instruments Z is associated with the common confounders U and thus proxies W. Both of these interpretations, however, are different ways to phrase the same problem. Instruments are not strongly relevant, as required by assumption 6.1.

With assumption 6.1, it is possible to perform the first debiasing step and construct a first step debiased moment as

$$m_1(O; k, \tau, g, q_k) = m_0(O; k) + q_k(Z)\left(g(Z) - \tau(Z) - k(A; \tau, g)\right) \tag{6.24}$$

$$\mathcal{Q}_k := \left\{q_k \in \mathcal{L}_2(Z) : q_k(Z) = \mathbb{E}\left[\xi_k(A)|Z\right] \ \forall \xi_k \in \mathcal{K}\right\} \tag{6.25}$$

$$\mathcal{Q}_{k,0} := \left\{q_k \in \mathcal{L}_2(Z) : q_k(Z) = \mathbb{E}\left[\xi_{k,0}(A)|Z\right] \ \forall \xi_{k,0} \in \Xi_{k,0}\right\} \tag{6.26}$$

Having taken care of the debiasing by k, we direct our attention to debiasing by τ. The functional $\tau \mapsto \mathbb{E}\left[m_1(O; k, \tau, g, q_k)\right]$ is indeed straigtforwardly linear and continuous in τ. Any dependence on τ trough k cancels out in expectation since m_1 is already debiased with respect to k. Thus, m_1 only depends on τ via $-q_k(Z)\tau(Z)$. The Riesz representer of $\tau \mapsto \mathbb{E}\left[m_1(O; k, \tau, g, q_k)\right]$ is the same as the Riesz representer of the simpler functional $\tau \mapsto \mathbb{E}\left[q_k(Z)\tau(Z)\right]$. Consequently, the results in section 6.1 apply, such that

$$\mathbb{E}\left[q_k(Z)\tau(Z)\right] = \mathbb{E}\left[\alpha_\tau(Z; q_k)\tau(Z)\right]$$

for some $\alpha_\tau(q_k) \in \mathcal{N}^{\perp}\left(P_{Z,\mathcal{T}}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}}\right) = \mathcal{N}^{\perp}\left(P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)} P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}\right)$. $\mathbb{E}\left[q_k(Z)\tau(Z)\right]$ already appears to be in some sort of Riesz representer form. However, the smoothness of q_k is not restricted, while the smoothness of $\alpha_\tau(q_k)$ implied by theorem 4 is essential to achieve robustness with respect to misestimation in τ. As $\alpha_\tau(q_k)$ depends on q_k, let $\alpha_{\tau,0} := \alpha_\tau(q_{k,0})$.

21

The second step debiased moment is

$$m_2(O; k, \tau, g, q_k, q_\tau) = m_1(O; k, \tau, g, q_k) + q_\tau(W; q_k)(\tau(Z) - g(Z)) \tag{6.27}$$

$$\mathcal{Q}_\tau(q_k) := \left\{ q_\tau \in \mathcal{L}_2(W) : q_\tau = \mathbb{E}\left[\xi_\tau(Z; q_k)|W\right], \ \forall \xi_\tau(q_k) \in \mathcal{T}, q_k \in \mathcal{Q}_k \right\} \tag{6.28}$$

$$\mathcal{Q}_{\tau,0}(q_k) := \left\{ q_\tau \in \mathcal{L}_2(W) : q_\tau = \mathbb{E}\left[\xi_{\tau,0}(Z; q_k)|W\right], \ \forall \xi_{\tau,0}(q_{k,0}) \in \Xi_{\tau,0}(q_k), q_k \in \mathcal{Q}_k \right\} \tag{6.29}$$

$$\Xi_{\tau,0}(q_k) := \arg\min_{\xi_\tau(q_k) \in \mathcal{T}} \left(\frac{1}{2} \mathbb{E}\left[\mathbb{E}\left[\xi_\tau(Z; q_k)|W\right]^2\right] - \mathbb{E}\left[q_k(Z)\xi_\tau(Z; q_k)\right] \right). \tag{6.30}$$

The only remaining task is debiasing by g, which requires defining g_0 in some form. Typically, we are interested in some sorts of averages over the outcome variable Y, hence for some known function $g_Y \in \mathcal{L}_2(Y)$ let

$$g_0(Z) := \mathbb{E}\left[g_Y(Y)|Z\right]. \tag{6.31}$$

Again, the Riesz representer of $g \mapsto \mathbb{E}\left[m_2(O; k, \tau, g, q_k, q_\tau)\right]$ is the same as the Riesz representer of a simpler functional, in this case $g \mapsto \mathbb{E}\left[(q_k(Z) - q_\tau(W; q_k)) g(Z)\right]$. While this functional is not immediately in Riesz representer form, it only takes an inner expectation to show that $\mathbb{E}\left[(q_k(Z) - q_\tau(W; q_k)) g(Z)\right] = \mathbb{E}\left[\alpha_g(Z; q_k, q_\tau)g(Z)\right]$ with $\alpha_g(Z; q_k, q_\tau) = \mathbb{E}\left[q_k(Z) - q_\tau(W; q_k)|Z\right]$. Hence, the third step debiased moment is

$$m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g) = m_2(O; k, \tau, g, q_k, q_\tau) + \alpha_g(Z; q_k, q_\tau)(g_Y(Y) - g(Z)) \tag{6.32}$$

$$\Xi_{g,0}(q_k, q_\tau) := \arg\min_{\xi_g(q_k, q_\tau) \in \mathcal{L}_2(Z)} \left(\frac{1}{2} \mathbb{E}\left[\xi_g(Z; q_k, q_\tau)^2\right] - \mathbb{E}\left[(q_k(Z) - q_\tau(W; q_k)) \xi_g(Z; q_k, q_\tau)\right] \right). \tag{6.33}$$

The third step debiased moment can be used construct a doubly robust estimator. First, note that each of the moments m_j for $j \in \{1, 2, 3\}$ identify θ_0.

Lemma 4. *Assume $\theta_0 = \mathbb{E}\left[m_0(O; k_0(\tau_0, g_0))\right]$ and 6.1. Then,*

$$\theta_0 = \mathbb{E}\left[m_0(O; k_0)\right] = \mathbb{E}\left[m_1(O; k_0, \tau_0, g_0, q_{k,0})\right]$$
$$= \mathbb{E}\left[m_2(O; k_0, \tau_0, g_0, q_{k,0}, q_{\tau,0})\right]$$
$$= \mathbb{E}\left[m_3(O; k_0, \tau_0, g_0, q_{k,0}, q_{\tau,0}, \alpha_{g,0})\right].$$

Each of m_j for $j \in \{1, 2, 3\}$ identify θ_0 because in each consecutive debiasing step a conditionally mean-zero term is added to the previous moment.

More importantly, the final debiased moment m_3 allows for a robust error decomposition.

Theorem 6 (Robust error decomposition). *Suppose $\theta_0 = \mathbb{E}\left[m_0(O; k_0(\tau_0, g_0))\right]$, assumption 6.1, and the conditions for theorem 4 hold for $q = q_\tau$. Then,*

$$\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g)\right] - \theta_0 = \mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\left(k(A; \tau, g) - k_0(A; \tau_0, g_0)\right)\right]$$
$$+ \mathbb{E}\left[\left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right)\left(\tau_0(Z; g_0) - \tau(Z; g)\right)\right]$$
$$+ \mathbb{E}\left[\left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right)\left(g(Z) - g_0(Z)\right)\right].$$

This robust error decomposition admits double robustness in corollary 1 and Neyman orthogonality in corollary 2. Double robustness here means that θ_0 is still correctly estimated when *either* the nuisances (k, τ, g) *or* the debiasing functions q_k, q_τ, α_g, which are closely related to Riesz representers, are estimated correctly. Interestingly, the debiasing functions depend on each other in opposite order compared to the ordinary nuisances. Specifically, while nuisance k depends on (g, τ), and nuisance τ depends on g, the debiasing function (Riesz representer) α_g depends on (q_k, q_τ), and the debiasing function q_τ depends on q_k.

Corollary 1 (Double robustness). *Suppose $\theta_0 = \mathbb{E}\left[m_0(O; k_0(\tau_0, g_0))\right]$, assumption 6.1, and the conditions for theorem 4 hold for $q = q_\tau$. Then, if*
$$\left((k = k_0(\tau_0, g_0)) \vee (q_k = q_{k,0})\right) \wedge \left((\tau = \tau_0(g_0)) \vee (q_\tau = q_{\tau,0})\right) \wedge \left((g = g_0) \vee (\alpha_g = \alpha_{g,0})\right),$$

$$\theta_0 = \mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g)\right].$$

A special form of double robustness and Neyman orthogonality holds in the sequential debiasing setting, where at each debiasing step, either the ordinary nuisance or the debiasing function must be correct, but never both. For example, when $\tau \notin \mathcal{T}_{\text{valid}}$, $q_\tau \in \mathcal{Q}_{\tau,0}(q_k)$ is necessary to still correctly estimate θ_0. However, it suffices that the ordinary nuisance is correct with $k \in \mathcal{K}_0$, while the debiasing function at the same stage may be incorrect as $q_k \notin \mathcal{Q}_{k,0}$. Specifically, double robustness and Neyman orthogonality is not restricted to all ordinary nuisances being jointly correct when a debiasing function is incorrect. Instead, the corresponding ordinary nuisance for the incorrect debiasing function must be correct. It suffices if at all other steps in the sequential debiasing that a given nuisance function affects either the ordinary nuisance or debiasing function is correct. This logic is captured by the conditions in corollary 2.

Corollary 2 (Neyman orthogonality). *Suppose $\theta_0 = \mathbb{E}\left[m_0(O; k_0(\tau_0, g_0))\right]$, assumption 6.1, and the conditions for theorem 4 hold for $q = q_\tau$. Let $k \in \mathcal{K}$, $\tau \in \mathcal{T}$, $g \in \mathcal{G}$, $k_0 \in \mathcal{K}_0$,*

$\tau_0 \in \mathcal{T}_{valid}$, $g_0(Z) = \mathbb{E}\left[g_Y(Y)|Z\right]$, and $q_k \in \mathcal{Q}_k$, $q_\tau \in \mathcal{Q}_\tau(q_k)$, $\alpha_g \in \mathcal{L}_2(Z)$, $q_{k,0} \in \mathcal{Q}_{k,0}$, $q_{\tau,0} \in \mathcal{Q}_{\tau,0}(q_k)$, $\alpha_{g,0} \in \Xi_{g,0}(q_k, q_\tau)$. Let $\partial_g f(X; g, h) := \frac{\partial}{\partial t} f(X; g_0 + tg, h)\big|_{t=0}$ for any functions f, g, and h. Then:

Under no further conditions,

$$0 = \frac{\partial}{\partial t} \mathbb{E}\left[m_3(O; k_0 + tk, \tau, g, q_k, q_{\tau,0}, \alpha_g)\right]\bigg|_{t=0} = \left(q_{k,0}(Z) - q_k(Z)\right).$$

If $\left((\partial_\tau k(\tau, g) = 0) \vee \left(q_k = q_{k,0}\right)\right)$,

$$0 = \frac{\partial}{\partial t} \mathbb{E}\left[m_3(O; k, \tau_0 + t\tau, g, q_k, q_{\tau,0}, \alpha_g)\right]\bigg|_{t=0}.$$

If $\left((\partial_g k(\tau, g) = 0) \vee (q_k = q_{k,0})\right) \wedge \left((\partial_g \tau(g) = 0) \vee (q_\tau = q_{\tau,0})\right)$,

$$0 = \frac{\partial}{\partial t} \mathbb{E}\left[m_3(O; k, \tau, g_0 + tg, q_k, q_\tau, \alpha_{g,0})\right]\bigg|_{t=0}.$$

If $\left((\tau = \tau_0(g_0)) \vee (q_\tau = q_{\tau,0})\right) \wedge \left((g = g_0) \vee (\alpha_g = \alpha_{g,0})\right)$,

$$0 = \frac{\partial}{\partial t} \mathbb{E}\left[m_3(O; k_0(\tau_0, g_0), \tau, g, q_{k,0} + tq_k, q_\tau, \alpha_g)\right]\bigg|_{t=0}.$$

If $\left((g = g_0) \vee (\alpha_g = \alpha_{g,0})\right)$,

$$0 = \frac{\partial}{\partial t} \mathbb{E}\left[m_3(O; k, \tau_0(g_0), g, q_k, q_{\tau,0} + tq_\tau, \alpha_g)\right]\bigg|_{t=0}.$$

Under no further conditions,

$$0 = \frac{\partial}{\partial t} \mathbb{E}\left[m_3(O; k, \tau, g_0, q_k, q_\tau, \alpha_{g,0} + t\alpha_g)\right]\bigg|_{t=0}.$$

Error bounds for estimation in corollary 3 explicitly reflect the sequential nature of debiasing function estimation. However, it is worth recalling that also the ordinary nuisances must be estimated sequentially, implicitly implying the same sequentiality in the opposite direction. As usual in debiased machine learning, error bounds are products of estimation errors from ordinary nuisance functions and debiasing functions. This product form of error bounds allows $\sqrt{n}$-estimation of θ_0 under the standard assumptions in debiased machine learning on faster than $n^{1/4}$-estimation of ordinary nuisances and debiasing functions and

appropriate sample splitting. Notably, when nuisances are defined via Fredholm integral equations as they are here, the error rates are generally rates of projections of nuisances. Crucially, having error bounds based on projections of nuisances eliminates the problem from ill-posedness [Bennett et al., 2022], which is a common feature of inverse problems.

Corollary 3 (Error bounds). *Suppose $\theta_0 = \mathbb{E}\left[m_0(O; k_0(\tau_0, g_0))\right]$, assumption 6.1, and the conditions for theorem 4 hold for $q = q_\tau$. Then,*

$$
\left| \mathbb{E}\left[m_3(O; \hat{k}_n, \hat{\tau}_n, \hat{g}, \hat{q}_{k,n}, \hat{q}_{\tau,n}, \hat{\alpha}_{g,n})\right] - \theta_0 \right|
$$

$$
\leq \min\left\{ \left\| P^{A,\mathcal{K}}_{\mathcal{L}_2(Z)}\left(k_0(\tau_0, g_0) - \hat{k}_n(\hat{\tau}_n, \hat{g}_n)\right) \right\|_2 \left\| q_{k,0} - \hat{q}_{k,n} \right\|_2, \ \left\| k_0(\tau_0, g_0) - \hat{k}_n(\hat{\tau}_n, \hat{g}_n) \right\|_2 \left\| P^{\mathcal{L}_2(Z)}_{A,\mathcal{K}}\left(q_{k,0} - \hat{q}_{k,n}\right) \right\|_2 \right\}
$$

$$
+ \min\left\{ \left\| P^{\mathcal{L}_2(W)}_{Z,\mathcal{T}}\left(\tau_0(g_0) - \hat{\tau}_n(\hat{g}_n)\right) \right\|_2 \left\| q_{\tau,0}(\hat{q}_{k,n}) - \hat{q}_{\tau,n}(\hat{q}_{k,n}) \right\|_2, \ \left\| \tau_0(g_0) - \hat{\tau}_n(\hat{g}_n) \right\|_2 \left\| P^{Z,\mathcal{T}}_{\mathcal{L}_2(W)}\left(q_{\tau,0}(\hat{q}_{k,n}) - \hat{q}_{\tau,n}(\hat{q}_{k,n})\right) \right\|_2 \right\}
$$

$$
+ \left\| (\hat{g}_n - g_0) \right\|_2 \left\| \alpha_{g,0}(Z; \hat{q}_{k,n}, \hat{q}_{\tau,n}) - \hat{\alpha}_{g,n}(Z; \hat{q}_{k,n}, \hat{q}_{\tau,n}) \right\|_2.
$$

Definition 6.3 (Debiased Machine Learning Estimator). Take K folds with index set $\mathcal{I}_j$ for $k = 1, 2, \ldots, K$ of (approximately) equal size. Construct all nuisance estimators K nuisance estimators $\eta^{(j)}$ with all data except data in $\mathcal{I}_j$ for $\eta \in \{k, \tau, g, q_k, q_\tau, \alpha_g\}$. The debiased machine learning estimator is

$$
\hat{\theta}_n = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} m_3(O; \hat{k}^{(j)}, \hat{\tau}^{(j)}, \hat{g}^{(j)}, \hat{q}_k^{(j)}, \hat{q}_\tau^{(j)}, \hat{\alpha}_g^{(j)}). \tag{6.34}
$$

Definition 6.4 (L2 Convergence Rates of Nuisance Estimators). Direct convergence rates:

$$
\left\| k_0(\tau_0, g_0) - \hat{k}_n(\hat{\tau}_n, \hat{g}_n) \right\|_2 \leq \epsilon_{n,k}
$$

$$
\left\| \tau_0(g_0) - \hat{\tau}_n(\hat{g}_n) \right\|_2 \leq \epsilon_{n,\tau}
$$

$$
\left\| \hat{g}_n - g_0 \right\|_2 \leq \epsilon_{n,g}
$$

$$
\left\| q_{k,0} - \hat{q}_{k,n} \right\|_2 \leq \epsilon_{n,q_k}
$$

$$
\left\| q_{\tau,0}(q_k) - \hat{q}_{\tau,n}(q_k) \right\|_2 \leq \epsilon_{n,q_\tau} \text{ for any } q_k \in \mathcal{Q}_k
$$

$$
\left\| \alpha_{g,0}(q_k, q_\tau) - \hat{\alpha}_{g,n}(q_k, q_\tau) \right\|_2 \leq \epsilon_{n,\alpha_g} \text{ for any } q_k \in \mathcal{Q}_k, q_\tau \in \mathcal{Q}_\tau
$$

Projected convergence rates:

$$\left\| P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}} \left(k_0(\tau_0, g_0) - \hat{k}_n(\hat{\tau}_n, \hat{g}_n) \right) \right\|_2 \leq \epsilon_{n,k,\mathcal{L}_2(Z)}$$

$$\left\| P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)} \left(q_{k,0} - \hat{q}_{k,n} \right) \right\|_2 \leq \epsilon_{n,q_k,\mathcal{K}}$$

$$\left\| P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}} \left(\tau_0(g_0) - \hat{\tau}_n(\hat{g}_n) \right) \right\|_2 \leq \epsilon_{n,\tau,\mathcal{L}_2(W)}$$

$$\left\| P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)} \left(q_{\tau,0}(q_k) - \hat{q}_{\tau,n}(q_k) \right) \right\|_2 \leq \epsilon_{n,q_\tau,\mathcal{T}} \text{ for any } q_k \in \mathcal{Q}_k$$

Assumption 6.2 (Rate conditions on nuisance estimators).   *1.* $\|\hat{\eta}_n - \eta_0\|_2^2 = o_p(1)$ *for any* $\eta \in \left\{ k, \tau, g, q_k, q_\tau, \alpha_g \right\}$.

2. The following requirements on the convergence rates in L2 norm hold:

$$o_p\left(n^{-1/2}\right) = \min\left\{ \epsilon_{n,k,\mathcal{L}_2(Z)} \epsilon_{n,q_k}, \epsilon_{n,k} \epsilon_{n,q_k,\mathcal{K}} \right\}$$

$$o_p\left(n^{-1/2}\right) = \min\left\{ \epsilon_{n,\tau,\mathcal{L}_2(W)} \epsilon_{n,q_\tau}, \epsilon_{n,\tau} \epsilon_{n,q_\tau,\mathcal{T}} \right\}$$

$$o_p\left(n^{-1/2}\right) = \epsilon_{n,g} \epsilon_{n,\alpha_g}$$

Assumption 6.2 is useful, because it can circumvent problems with possibly poor convergence rates in ill-posed inverse problems as long as the projections of nuisance functions are estimated with sufficient convergence rates, which may be much easier to satisfy. Firstly, we require mean square convergence in probability of each nuisance function estimator. In a more general framework, [Bennett et al., 2022] in assumptions 12-21b provide extensive conditions under which mean square convergence in probability is satisfied for minimum-norm nuisance functions identified as solutions to Fredholm integral equations of the first kind. The requirements on convergence rates in L2 norm in assumption 6.2 are doubly robust, and allow for convergence rates of projections instead of the nuisance functions themselves. As a result, estimation is robust not only in the typical debiased machine learning sense to slower than parametric estimation of nuisance parameters, but also to the ill-posedness of inverse problems that define some nuisances by only requiring L2 norm convergence in more favourable projections. For example, it suffices if all projections of nuisances identified by integral equations satisfy $o_p\left(n^{-1/4}\right)$ convergence in L2 norm.

Theorem 7 (Debiased Estimation of Target Parameter). *Let $\hat{\theta}_n$ be a debiased machine learning estimator as in definition 6.3. Suppose $\theta_0 = \mathbb{E}\left[m_0(O; k_0(\tau_0, g_0))\right]$, assumption 6.1,*

assumption 6.2, and the conditions for theorem 4 hold for $q = q_\tau$. Then,

$$\hat{\theta}_k - \theta_0 = \mathbb{E}_{n,k}\left[m_3(O; k_0, \tau_0, g_0, q_{k,0}, q_{\tau,0}, \alpha_{g,0}) - \theta_0\right] + o_p\left(n^{-1/2}\right)$$

$$\sqrt{n}\left(\hat{\theta}_k - \theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(m_3(O; k_0, \tau_0, g_0, q_{k,0}, q_{\tau,0}, \alpha_{g,0}) - \theta_0\right) + o_p(1)$$

$$\to N(0, \sigma_0^2), \quad \sigma_0^2 = \mathbb{E}\left[\left(m_3(O; k_0, \tau_0, g_0, q_{k,0}, q_{\tau,0}, \alpha_{g,0}) - \theta_0\right)^2\right]$$

$$\frac{\sqrt{n}}{\hat{\sigma}_n^2}\left(\hat{\theta}_k - \theta_0\right) \to N(0,1) \ \ for \ any \ \ \hat{\sigma} \ \ s.t. \ \ \hat{\sigma}_n^2 \to \sigma_0^2 \ \ in \ probability.$$

7   Practical Guide

Straightforward testing and discussion of model assumptions is key in any application. In this section, we provide a practical guide to identification with this approach. Each of the four steps we describe has its own subsection. As in standard IV, the relevance assumptions generally remain testable. The conditional exogeneity assumption is not testable up to a specification test.

7.1   Find T and test relevance of Z, W for U

In this step, we test the relevance of W for U (assumption 2.1.3), as well as the relevance of Z for U. The latter is a necessary condition for the relevance of Z for A conditional on T (assumption 2.1.2c), which is tested explicitly in subsection 7.2.

First, we find some $T = \tau(Z)$ such that $Z \perp\!\!\!\perp W \mid T$ is satisfied. As long as some $T = \tau(Z)$ leads to the conditional independence of Z and W, this control function $\tau \in \mathcal{L}_2(Z)$ also renders Z and U conditionally independent. To test for the sufficient relevance of Z and W with respect to U, we can use T. A sufficient condition for relevance of Z and W for U is that both Z and W contain spare information conditional on T. This motivates the choice of a valid $\tau \in \mathcal{T}_{\text{valid}}$, which captures the least information about Z while ensuring the conditional exogeneity of Z conditional on T, as discussed in section 3. To simplify the argument, suppose that our model is linear and the dimensions for (Z, T, W, U) are (d_Z, d_T, d_W, d_U). Often, relevance of Z and W for U imply $\min\{d_Z, d_W\} \geq d_U$. The minimum dimension of T to ensure conditional independence of Z and W is d_U, so we know that any $T = \tau(Z)$ such that $Z \perp\!\!\!\perp W \mid T$ must satisfy $d_T \geq d_U$. If we can reject a test with the null hypothesis

$$H_0 : \min\{d_Z, d_W\} \leq d_T, \ \text{and alternative} \ H_a : \min\{d_Z, d_W\} > d_T,$$

this implies $\min\{d_Z, d_W\} > d_U$. Thus, there is a test whether Z and W are relevant for U.

However, $\min\{d_Z, d_W\} > d_T$ is a sufficient, not a necessary condition for the relevance of Z and W for U. Z and W are still relevant for U when $\min\{d_Z, d_W\} = d_U$. Unfortunately, this hypothesis is not testable with unobserved U. So, how should an applied researcher proceed when $\min\{d_Z, d_W\} = d_T$ (which could mean $\min\{d_Z, d_W\} = d_U$)? Here, we need to distinguish $d_Z = d_T$ from $d_W = d_T$.

$d_Z = d_T$ When T contains as much information as Z, there is no point in moving to step 2. The instruments Z contain no variation conditional on T, so Z cannot be relevant for treatment A conditional on T.

$d_W = d_T$ We know for sure that $d_W \leq d_U$. If $d_W = d_U$, the proxies W are exactly relevant for U without spare information. If we are willing to assume $d_W = d_U$, we could move forward to step 2. The variation in U associated with Z would still be held fixed with T in this case. Yet, $d_W = d_U$ is not testable. It may well be that $d_W < d_U$. Then, the variation in U associated with Z is *not* held fixed with T. We can never test the completeness of W for U when $d_W = d_T$. Accordingly, we do not generally suggest to move to step 2 by relying on the assumption $d_W = d_U$, when $d_W = d_T$ is observed.

7.2   Test relevance of Z for A given T

In step 1, the relevance of Z, W for U was confirmed. Now, we test the relevance of the instruments Z for treatment A conditional on T (assumption 2.1.2c).

Depending on the additional model assumptions, this is either a test of completeness (2.1.2c), or common support (4.3). As any $T = \tau(Z)$ is simply the control function $\tau \in \mathcal{L}_2(Z)$ applied to instruments Z, the test of relevance of Z for A given T is straightforward for any given τ. It is as simple as a test of instrument relevance with observed confounders. Let us return to a linear model. If all components of the instrument vector Z are correlated with both U and A, the conditional relevance requirement simplifies to $(d_Z - d_T) \geq d_A$. After conditioning on all variation in Z which correlates with U by holding T fixed, $d_Z - d_T$ dimensions of instruments Z remain to infer the causal effect of treatment A on outcome Y. The remaining instrument variation of dimension $d_Z - d_T$ is relevant for treatment A only if the treatment's dimension d_A is smaller than or equal to $d_Z - d_T$.

If Z is found to be relevant for A given T, it implies that Z is relevant for both U and A. Interpreting U as any source of unobserved variation which associates instruments Z and proxies W, Z would have no variation conditional on T if they were not sufficiently relevant for U ($d_Z < d_U$). So, if Z is relevant for A given T, it implies that Z already had to be relevant for U.

7.3   Exogeneity of Z conditional on U

In step 1 and 2 we tested all relevance assumptions in this model. The conditional exogeneity assumption for Z conditional on U remains untestable. To be precise, $Y \perp\!\!\!\perp Z \mid (A, U)$ (assumption 2.1.2a) can only be justified on theoretic grounds, not observed data. In order to justify conditional exogeneity theoretically, T can be used to understand the unobserved common confounders U. As T captures all variation in Z associated with U, T immediately explains U in terms of its association with Z and W. From the association of T and W, we can interpret U even better. For example: If subject-specific pre-college GPA measures are used as instruments Z, and T turns out to capture average GPA, then U could be interpreted as general ability. Suppose W contains dummies capturing whether someone has engaged in risky behaviour, including drugs and illegal activity, while in high school. From theory and empirical evidence we would expect high ability to lead to less risky behaviour. Thus, if T is an average GPA, it would be expected to negatively correlated with W.

Once we used T to understand the variation reflected by the unobserved confounders U, we can construct a theoretic argument with respect to the conditional exogeneity of instruments Z. In our example, the common confounder U reflected general ability. The conditional exogeneity assumption reduces to whether conditional on general ability U, the subject-specific pre-college GPA measures Z are excluded. This argument clearly depends on the respective choice of treatment A and outcome Y.

As in standard IV, specification tests, which can be revealing about the exogeneity of Z, are possible if the model is overidentified. If a specification test suggests that different subset of instruments Z conditional on T result in estimators with different probability limits, we reject that all instruments Z are excluded conditional on T (unless the estimand is the local average treatment effect which can vary across subpopulations). Necessary for any such test is model overidentification for the causal effect θ_0 conditional on T. In a simple linear model, overidentification would e.g. mean $(d_Z - d_T) > d_A$. After keeping d_T dimensions of Z fixed, the instruments must still contain overidentifying information.

Ultimately, just as in standard IV, the conditional exogeneity assumption for Z remains largely untestable. Therefore, it is crucial to better understand U from the control variable T.

7.4   Estimation

In the final fourth step, use Z to instrument for treatment A conditional on control variable T, to identify (and estimate) the structural function or average causal effect θ_0 of A on Y. Having established that all necessary relevance and exogeneity requirements hold, an

estimator $\hat{\theta}$ can be formulated for the causal effect of interest θ_0. The form of this estimator depends on the type of parametric model assumptions made.

8   Example: Linear Returns to Education

Interested in the returns to education, we use data from the National Longitudinal Survey of Youth 1997 [Bureau of Labor Statistics, 2019]. The variables of interest are introduced below.

Y Household net worth at 35: continuous variable, in USD.

A BA degree: 1 if individual i obtained a BA degree, 0 otherwise.

Z Pre-college test results: subject-specific and overall GPA; ASVAB percentile.

W Risky behaviour dummies: whether i drank, smoked, or engaged in other behaviours considered risky by the age of 17.

U Ability: Unmeasured intellectual capacity.

$\tilde{U}$ Other biases: Selection on unobservables into obtaining a BA degree (at least in part result of optimising individuals).

X Covariates: sex, college GPA, parental education/net worth, siblings, region, etc.

A review of the vast literature on returns to education is far beyond the scope of this paper [Psacharopoulos and Patrinos, 2018]. Instead, we focus on estimation of a very specific return to education: The causal effect of obtaining a bachelor's degree A on household net worth at 35. Even in a simple linear model like

$$Y = \alpha_Y + A\beta + U\gamma_Y + W\upsilon_Y + X\eta_Y + \varepsilon_Y, \tag{8.1}$$

two distinct potential sources of confounding are easily identified via the unobservable components of 8.1:

$U\gamma_Y$ Ability U likely has a positive effect on household net worth Y, by means of salary and non-salary based net worth accumulation [Griliches, 1977]. The vector-valued linear parameter γ_Y captures this positive linear effect of ability on net worth.

ε_Y The disturbance ε_Y captures all variation in Y, which is jointly unexplained by (A, U, W, X). This can be understood as individual-specific, heterogeneous characteristics, and chance.

Any correlation of A with either of these terms leads to biased estimates of β.

How does obtaining a BA degree A correlate with ability U and a general disturbance ε_Y?

In this identification problem, selection bias in inherent. At least to some degree, individuals choose whether to obtain a BA degree as a result of an optimisation problem of expected utility subject to an information set $\mathcal{I}$:

$$A = \arg\max_{a \in \{0,1\}} \left(\mathbb{E}\left[u(Y(a)) - c(a) | A = a, \mathcal{I}\right] \right), \tag{8.2}$$

where $u : \mathcal{Y} \to \mathbb{R}$ is a utility function for net worth with diminishing returns, and $c : \{0,1\} \to \mathbb{R}$ a cost function for obtaining a BA degree A. Both utility and cost function can be individual-specific.

For ease of illustration, suppose individuals are perfectly informed with $\mathcal{I} = (A, U, W, X, \varepsilon_Y)$. Then, each individual chooses $a \in \{0,1\}$ to maximise the utility associated with potential outcome $Y(a)$ minus cost $c(a)$. In this case, there is an easy decision rule to determine optimal A:

$$A = \arg\max_{a \in \{0,1\}} u(Y(a)) - c(a),$$
$$= \mathbb{1}\left[u(Y(1)) - u(Y(0)) > c(1) - c(0)\right].$$

$U\gamma_Y$ Ceteris paribus, an increase in ability U equally increases $Y(0)$ and $Y(1)$ according to model 8.1. Due to diminishing returns in the utility function u, $u(Y(1)) - u(Y(0))$ decreases. However, also $c(1)$ decreases as higher-ability individuals experience a lower utility cost of obtaining a BA degree. The overall effect on the choice of A is ambiguous and depends on the utility functions of the individual.

ε_Y The effect of ε_Y on the choice of A, on the other hand, is unambiguous. An increase in ε_Y reduces $u(Y(1)) - u(Y(0))$ due to the diminishing returns of u. Cost c, however, is unaffected by ε_Y. Hence, A inevitably negatively correlates with ε_Y.

This logic regarding negative selection bias when treatment is chosen by utility-maximising individuals is by no means novel [Heckman et al., 2006], or unique to the returns to education identification problem. Negative selection bias is inherent to the treatment variable when it is at least in part the result of optimising behaviour by utility-maximising heterogeneous individuals. Novel in our approach is the ability to explicitly account for certain biases, in this case ability bias, when proxies for them exist. Finding excluded instruments can be much more straightforward when pertinent biases, like ability bias, have already been taken care of.

In our identification approach, instruments Z are pre-college test results. These results are strongly correlated with ability U. Yet, conditional on ability, and some other covariates, pre-college test results contain random variation, which is excluded with respect to household net worth Y (at age 35). Concurrently, even random variation in pre-college test results is a strong predictor of obtaining a BA degree. Hence, instrument relevance likely holds. The proxies W are dummies for whether an individual engaged in risky behaviours at high school age. Among others, the risky behaviour dummies include drinking, smoking (marijuana), selling drugs and stealing. Theory and empirical evidence suggest the correlation of low intelligence and risky behaviour [Loeber et al., 2012]. Therefore, ability U both causes instruments Z and proxies W in our data. Ability U is the common confounder in this causal question. Clearly, additional covariates are necessary to justify instrument exogeneity. These include sex, college GPA, parental education and net worth, the number of siblings, region of residence, etc.

8.1 Assumptions

The linear equivalent to the general common confounding IV model in assumption 2.1 is described as assumption 8.1. Again, for ease of notation assume $d_A = 1$, just as in this returns to education identification problem.

Assumption 8.1 (Linear IV Model with Common Confounding)**.**

1. *Linear outcome model projection:*

$$Y = \alpha_Y + A\beta + U\gamma_Y + W\upsilon_Y + X\eta_Y + \varepsilon_Y \tag{8.3}$$

2. *Instruments*

   (a) *Exogeneity:* $\mathbb{E}\left[\varepsilon_Y(Z, U, W, X)\right] = \mathbf{0}$.

   (b) *Relevance: For the linear projection of A on (Z, U, W, X),*

   $$A = \alpha_A + Z\zeta + U\gamma_A + W\upsilon_A + X\eta_A + \varepsilon_A, \ \mathbb{E}\left[\varepsilon_A(Z, U, W, Z)\right] = \mathbf{0} \tag{8.4}$$
   $$\mathrm{rank}\left(\mathbb{E}\left[(Z\zeta)\, A \mid T, X\right]\right) = d_A. \tag{8.5}$$

3. *Proxies*

(a) Exogeneity: For the linear projection of W on (Z, U, X) and (Z, X),

$$W = \alpha_W + U\gamma_W + X\eta_W + \varepsilon_W, \qquad \mathbb{E}\left[\varepsilon_W(Z, U, X)\right] = \mathbf{0}, \qquad (8.6)$$

$$W = \tilde{\alpha}_W + Z\tilde{\gamma}_W + X\tilde{\eta}_W + \tilde{\varepsilon}_W, \qquad \mathbb{E}\left[\tilde{\varepsilon}_W(Z, X)\right] = \mathbf{0}. \qquad (8.7)$$

with $T := Z\tilde{\gamma}_W + X\tilde{\eta}_W$.

(b) Relevance: $\operatorname{rank}(\gamma_W) \geq d_U$

$$. \qquad (8.8)$$

To simplify notation, let $Z_{|X}$ be the true residual of a projection of Z onto X. The linearity of the outcome model implies that the covariance $\operatorname{Cov}\left[(Z_{|X}\zeta), Y\right]$ is

$$\operatorname{Cov}\left[(Z_{|X}\zeta), Y\right] = \operatorname{Cov}\left[(Z_{|X}\zeta), A\right]\beta + \operatorname{Cov}\left[(Z_{|X}\zeta), U\right]\gamma_Y + \operatorname{Cov}\left[(Z_{|X}\zeta), W\right]\upsilon_Y.$$

The above expression uses the uncorrelatedness of Z and ε_Y in assumption 8.1.2a. If it were not for the linear confounding from the unobserved common confounders U and proxies W, Z would be excluded. Next, we demonstrate how to use the proxies W to keep $\operatorname{Cov}\left[(Z\zeta)U\right]$ fixed.

$$\operatorname{Cov}\left[(Z_{|X}\zeta), U\right] = \operatorname{Cov}\left[(Z_{|X}\zeta), W\right]\gamma_W^{\mathsf{T}}\left(\gamma_W\gamma_W^{\mathsf{T}}\right)^{-1}$$

The inverse $\left(\gamma_W\gamma_W^{\mathsf{T}}\right)^{-1}$ exists under assumption 8.1.3 that the rank of γ_W is at least d_U. Then, slightly rewriting $\operatorname{Cov}\left[(Z\zeta)Y\right]$ as

$$\operatorname{Cov}\left[(Z_{|X}\zeta), Y\right] = \operatorname{Var}\left[Z_{|X}\zeta\right]\beta + \operatorname{Cov}\left[(Z_{|X}\zeta), W\right]\tilde{\upsilon}_W,$$

$$\tilde{\upsilon}_Y := \upsilon_Y + \upsilon_A\beta + \gamma_W^{\mathsf{T}}\left(\gamma_W\gamma_W^{\mathsf{T}}\right)^{-1}(\gamma_Y + \gamma_A\beta),$$

implies that any endogeneity of residualised instruments $Z_{|X}$ is controlled for by conditioning on $Z\tilde{\gamma}_W$ from the linear projection 8.7. To be precise,

$$\operatorname{Cov}\left[(Z_{|X}\zeta), Y | Z\tilde{\gamma}_W\right] = \operatorname{Var}\left[Z_{|X}\zeta | Z\tilde{\gamma}_W\right]\beta + \underbrace{\operatorname{Cov}\left[(Z_{|X}\zeta)W | Z\tilde{\gamma}_W\right]}_{=0}\tilde{\upsilon}_Y.$$

The covariance of the first stage can be rewritten as

$$\text{Cov}\left[(Z_{|X}\zeta), A|Z\tilde{\gamma}_W\right] = \text{Var}\left[Z_{|X}\zeta|Z\tilde{\gamma}_W\right] + \underbrace{\text{Cov}\left[(Z_{|X}\zeta)W|Z\tilde{\gamma}_W\right]}_{=0}\tilde{v}_A.$$

Using both of these results, and one-dimensional treatment A to simplify notation, a simple ratio form for the linear effect of A on outcome Y is

$$\beta = \frac{\text{Cov}\left[\left(Z_{|X}\zeta\right)Y \mid Z\tilde{\gamma}_W\right]}{\text{Cov}\left[\left(Z_{|X}\zeta\right)A \mid Z\tilde{\gamma}_W\right]} = \frac{\text{Cov}\left[(Z\zeta)Y \mid T, X\right]}{\text{Cov}\left[(Z\zeta)A \mid T, X\right]}, \text{ with } T = Z\tilde{\gamma}_W. \tag{8.9}$$

Hence, the estimator differs from standard IV based estimation only by also holding a linear prediction T of W fixed as the partial predicted values $Z\zeta$ for A change. Thus, the relevance requirement 8.1.2b for the instruments Z is conditional on T and X. T can be represented by a d_U-dimensional linear function of Z, $\mathbb{E}[U|Z, X]$, multiplied by γ_W. Hence, a simpler way to understand the relevance requirement 8.1.2b is as

$$\text{rank}(\zeta) \geq (d_U + d_A) \tag{8.10}$$

A total of d_U dimensions of variation in Z are typically needed to account for the d_U-dimensional confounding effect of U via $\mathbb{E}[U|Z, X]$, while the remaining variation in Z still needs to be relevant for A. Other than in trivial cases[2], equation 8.10 describes this relevance requirement satisfactorily as a rank condition on ζ, the partial linear projection effect of Z on A conditional on (U, W, X).

8.2   Find T and test relevance of Z, W for U

A valid control function is the linear prediction $T = Z\tilde{\gamma}_W$ under assumption 8.1.3, meaning that conditional on (T, X), instruments Z are still relevant for A. However, its OLS estimate $T = Z\hat{\tilde{\gamma}}_W$ generally is not a valid control function, because T and Z are perfectly correlated due to sampling variation, unless $d_Z > d_W$. However, even when $d_Z > d_W$, the true $\tilde{\gamma}_W$ will have rank $d_U \leq d_W$, while its OLS estimate $\hat{\tilde{\gamma}}_W$ always has possibly larger than necessary rank d_W due to sampling variation. Ultimately, the estimate $\hat{\tilde{\gamma}}_W$ should at best have exactly rank d_U. A test is needed for the rank r_0 of matrix $\tilde{\gamma}_W$. If $\mathbb{E}[U|Z, X] = Z\gamma_Z + X\gamma_X$, then $\tilde{\gamma}_W = \gamma_Z\gamma_W$. Sufficient for the rank condition in assumption 8.1.3 is $r_0 < \min\{d_Z, d_W\}$. This condition means that an unobservable variable of smaller dimension than both W and Z can

---

[2]e.g. when Z contains perfectly collinear variation conditional on (U, W, X).

explain all correlation between W and Z conditional on X. This unobserved variable is the common confounder U. By the definition of U as the (minimum information) unobserved variable which renders W and Z mean-independent, γ_Z has $d_U \leq d_Z$ linearly independent rows ($\mathrm{rank}(\gamma_Z) = d_U$). As γ_W has dimensions $d_U \times d_W$ and $d_U \leq d_W$, $\mathrm{rank}(\gamma_W) \leq d_U$ and thus $r_0 = \mathrm{rank}(\gamma_Z \gamma_W) = \mathrm{rank}(\gamma_W)$.

While $r_0 < d_W$ suffices to confirm the relevance of W for U in assumption 8.1.3, $r_0 < d_Z$ is necessary for Z to be relevant for treatment A in assumption 8.1.2b. A suitable test for some $r < \min\{d_Z, d_W\}$ has null hypothesis

$$H_0 : r_0 \leq r, \text{ and alternative } H_a : r_0 > r. \tag{8.11}$$

With the OLS estimator $\hat{\tilde{\gamma}}_W$, we apply a bootstrap based test for its rank. First, write the singular value decomposition as

$$\tilde{\gamma}_W = \underset{d_Z \times d_Z}{P_0} \underset{d_Z \times d_W}{\Pi_0} \underset{d_W \times d_W}{Q_0^\intercal} \tag{8.12}$$

Then, let $\phi_r(A) := \sum_{j=r+1}^{m_A} \pi_j^2(A)$ be the sum of squared singular values of A from the $(r+1)$ largest to the smallest singular value, which is the m_A-th singular value, where m_A is the minimum across A's number of rows and columns. Then, an equivalent test to 8.11 is a test with null hypothesis

$$H_0 : \phi_r\left(\tilde{\gamma}_W\right) = 0, \text{ and alternative } H_a : \phi_r\left(\tilde{\gamma}_W\right) > 0. \tag{8.13}$$

The bootstrap procedure is as follows:

1. For each binary proxy $W_j \in W$, calculate the probability $p_j := \mathrm{Pr}\left(W_j = 1\right)$ under H_0 as $p_{j,0} = \mathrm{Logit}\left(\left(\underset{d_Z \times r}{Z P_{0,r}} \underset{r \times r}{\Pi_{0,r}} \underset{r \times 1}{Q_{0,r,j}^\intercal} + \underset{d_X \times 1}{X \tilde{\eta}_{W,j}}\right)\beta_0 + \alpha_0\right)$, where

   (a) $P_{0,r}$ corresponds to the first r columns of P_0,

   (b) $Q_{0,j}$ corresponds to the first r entries of the j-th row of Q_0

   (c) $\Pi_{0,r}$ corresponds to the $r \times r$ matrix of of the first r rows and columns of Π_0,

   (d) $\tilde{\eta}_{W,j}$ corresponds to j-th column of $\tilde{\eta}_W$,

   (e) β_0 and α_0 are univariate coefficients, which need to be estimated.

2. Draw 1000 new bootstrap samples $b \in \mathcal{B}$ of binary proxies as W_0^b using the $n \times d_W$ probability matrix $(p_{0,0}, p_{1,0}, \ldots, p_{d_W,0})$.

3. For each bootstrap sample $b \in \mathcal{B}$: Calculate the sample projection coefficient $\hat{\tilde{\gamma}}_{W,0}^b$ by projecting W_0^b onto (Z, X) (all demeaned), and the sum of its smallest squared singular values starting at the $(r + 1)$ largest as $\phi_{r,0,b} := \phi_r\left(\hat{\tilde{\gamma}}_{W,0}^b\right)$.

4. Obtain the p-value as $1 - \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \mathbb{1}\left(\phi_{r,0,b} < \phi_r(\tilde{\gamma}_W)\right)$.

Figure 2: Bootstrap based test for $H_0 : \text{rank}(\tilde{\gamma}_W) = r_0 \leq r$
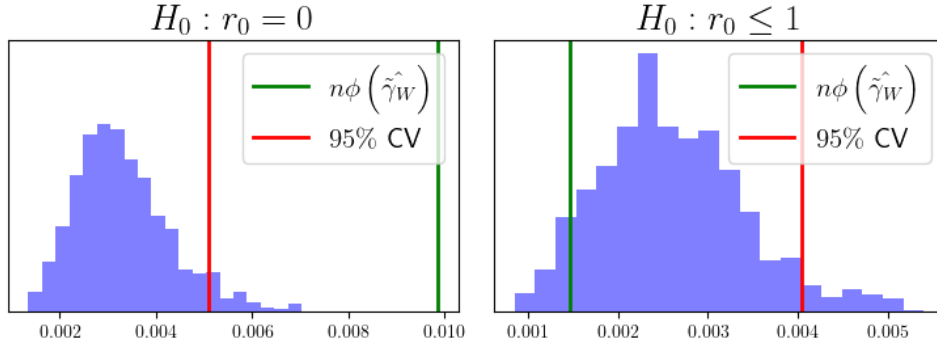
Notes: This figure illustrates the bootstrap distribution of the test statistic $n\phi_r\left(\hat{\tilde{\gamma}}_W\right)$ in the test with null hypothesis $H_0 : \text{rank}(\tilde{\gamma}_W) = r_0 \leq r$. The left figure depicts the test statistic distribution under $r = 0$, and the left under $r = 1$. For $r = 0$, the p-value is zero. $H_0 : r_0 = 0$ is strongly rejected. For $r = 1$, the p-value is 0.933. $H_0 : r_0 \leq 1$ cannot be rejected at any meaningful level of significance.

In figure 2, the bootstrapped distributions of the test statistic $n\phi_r\left(\hat{\tilde{\gamma}}_W\right)$ are depicted under two different null hypotheses: $r_0 = 0$ and $r_0 \leq 1$. Non-rejection of the test is evidence in favour of the low rank r_0 of $\tilde{\gamma}_W$. In the left diagram of figure 2, where the test concerns $r_0 = 0$, the p-value is at zero. The test provides strong evidence against $r_0 = 0$, which indicates some correlation between W and Z conditional on X. The right diagram of figure 2 depicts the test statistic bootstrap distribution for $H_0 : r_0 \leq 1$, and provides strong evidence against rejection. The associated p-value is 93.3%. Thus, we can conclude that the rank of γ_W is at most one. In the NLS97 data, pre-college test results Z and risky behaviour dummies have dimensions $d_Z = 7$ and $d_W = 9$. Thus, $r_0 \leq 1$ allows the conclusion that the common confounder dimension is small: $d_U \leq 1$. Conditional on covariates X, all covariance between Z and W is explained by a one-dimensional unobserved U. Successfully, the proximal assumption 8.1.3 was tested. In addition, the necessary $d_Z > d_U$ condition for conditional instrument relevance (assumption 8.1.2b) was confirmed.

8.3 Test relevance of Z for A given T

Despite satisfying the necessary $d_Z > d_U$ condition for IV relevance (assumption 8.1.2b), a proper test for the conditional relevance of Z for A given the control function T is still missing. In this step, we first explain how to construct the here one-dimensional control variable T after having conducted the tests in section 8.3. Then, we test for the conditional relevance of instrument Z for treatment A given this control function T.

Given the statistical evidence in favour of $d_U \leq 1$, we construct the variable

$$\underset{N\times1}{T} := \underset{N\times d_Z}{Z} \underset{d_Z\times1}{\hat{P}_{0,1}} \underset{1\times1}{\hat{\Pi}_{0,1}}, \tag{8.14}$$

with the singular value decomposition of the OLS estimator $\hat{\tilde{\gamma}}_W = \hat{P}_0\hat{\Pi}_0\hat{Q}_0^\intercal$. $\hat{P}_{0,1}$ is the first column of $\hat{P}_0$, and $\hat{\Pi}_{0,1}$ is the top-left entry of $\hat{\Pi}_0$. Aside from sampling error, proxies W are mean-independent from instruments Z conditional on (T, X).

With the control T now defined, we can use a bootstrap based test to confirm the relevance of instruments Z for A conditional on (T, X). The null hypothesis can be formulated as

$$H_0 : \text{rank}\left(\mathbb{E}\left[(Z\zeta)A|T, X\right]\right) < d_A, \text{ with alternative } H_a : \text{rank}\left(\mathbb{E}\left[(Z\zeta)A|T, X\right]\right) = d_A. \tag{8.15}$$

Importantly, under H_0 the effect of Z on A (given X) would be fully described by a one-dimensional T, as the dimension of U was found to be $r_0 \leq 1$ in section 8.2. When treatment A is one-dimensional, a simple test for this null hypothesis compares the R^2 of an unrestricted (8.16) and restricted regression (8.17).

$$A = \tilde{\alpha}_{A,ur} + Z\tilde{\zeta} + X\tilde{\eta}_{A,ur} + \tilde{\varepsilon}_{A,ur}, \qquad \mathbb{E}\left[\tilde{\varepsilon}_{A,ur}|Z, X\right] = 0, \tag{8.16}$$
$$A = \tilde{\alpha}_{A,r} + T\tilde{\gamma}_A + X\tilde{\eta}_{A,r} + \tilde{\varepsilon}_{A,r}, \qquad \mathbb{E}\left[\tilde{\varepsilon}_{A,r}|T, X\right] = 0. \tag{8.17}$$

Under H_0, both regressions would predict A equally well, despite the dimension reduction on Z in the second regression, 8.17. With the uncertainty in estimated T, we use a simple bootstrap-based test. With 1000 bootstrap samples $b_t \in \mathcal{B}_t$, we obtain a bootstrap distribution of R^2. Under H_0, R^2_{ur} is (asymptotically) distributed as R^2_r.

Figure 3 depicts the bootstrap distribution of R^2_r based on the restricted regression 8.17. The control variable T is constructed for each bootstrap sample as described in 8.14. The unrestricted R^2_{ur} based on the unrestricted regression 8.16 fits the data significantly better, which indicates rejection of H_0. The p-value is 0.023. There is predictive information in Z for A, beyond that controlled for in (T, X). In other words, Z satisfies the conditional

Figure 3: Test for Relevance of Z for A given T $\left(H_0 : \mathrm{rank}\left(\mathbb{E}\left[Z\zeta|T,X\right]\right) < d_A\right)$
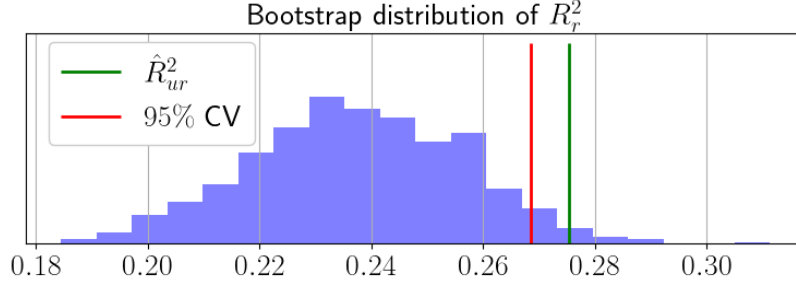
Notes: This figure illustrates the bootstrap distribution of the restricted R_r^2 in regression 8.17. The dimension of T,$d_T = 1$, is based on the test in section 8.3.

instrument relevance requirement 8.1.2b.

8.4 Exogeneity of Z conditional on U

While both relevance assumptions 8.1.3 and 8.1.2b could be tested successfully, the exogeneity of instrument Z conditional on U in assumption 8.1.2a remains generally untestable.

To argue whether Z is exogenous conditional on U, it is worth asking: Which information is being held fixed in T, and what does this imply about U? The linear construction of T from Z is illustrated in table 1, where the instruments have been normalised to standard deviation one. Z is mean-independent from W conditional on (T, X). T mostly consists of an average of subject-specific pre-college GPA measures. In this sense, T closely measures academic ability, as captured by pre-college GPA measures. Without the transcript GPA and ASVAB percentile, the subject-specific GPA measures describe 94.5% of variation in T. Despite the negative dependence of T on ASVAB percentile in its construction, T positively correlates with ASVAB percentile unconditional on the GPA measures with a 0.31 correlation coefficient. The interpretation of T and consequently U is pretty straightforward: It positively reflects (academic) ability.

Table 1: Construction of $T = Z\hat{P}_{0,1}\hat{\Pi}_{0,1}$

	GPA				ASVAB
	English	Math	SocSci	LifeSci	percentile
T	0.537	0.216	0.280	0.214	-0.262

As U reflects (academic) ability, an increase in T is expected to result in a reduction of risky behaviour [Loeber et al., 2012]. Indeed, a one standard deviation increase in T reduces

the probability of having engaged in risky behaviour by the age of 17 between 3% and 9%, as illustrated in table 2. All effects have strong statistical and economic significance. Compared to the average probability of engaging in risky behaviour, the estimated effect of a one standard deviation change in T is largest for some of the riskiest behaviour we considered: selling drugs (-54%), running away (-46%), and attacking someone (-41%).

Table 2: Effect of T on W

	drink	smoke	try marijuana	run away	attack someone	sell drugs	destroy property	steal < 50$	steal > 50$
Pr	65.5%	47.1%	29.8%	10.9%	19.0%	9.1%	32.0%	38.1%	8.0%
T	-7.9%	-9.5%	-9.6%	-5.0%	-7.8%	-4.9%	-3.5%	-4.7%	-3.2%

Notes: The table contains sample probabilities for engaging in risky behaviour by the age of 17 in the Pr row. The estimated decrease in the probability of engaging in risky behaviour from a linear probability model for a one standard deviation increase in T is noted in the row corresponding to T.

T captures the information we expected based on our suspicion about the unobserved confounder ability. T closely reflects (academic) ability as measured by high-school GPA measures, which reduces the probability of engaging in risky behaviours during high-school. Thus, we can conclude that the common confounder U contains the unobserved variable ability.

Now, an argument is required for the conditional exogeneity of instruments Z given unobserved ability U and observed covariates X:

$$Y = \alpha_Y + A\beta + U\gamma_Y + W\upsilon_Y + X\eta_Y + \varepsilon_Y, \qquad \mathbb{E}\left[\varepsilon_Y \mid Z, X\right] = 0.$$

While ability is the obvious confounder of the effect of pre-college GPA measures on net worth later in life, there are other possible confounders. Among everyone who goes to college, those with higher pre-college GPA are likely to also have a higher college GPA. Even conditional on whether someone obtained a BA degree, a higher college GPA likely leads to higher earnings later in life. Thus, college GPA is an important observed confounder. Family and individual net worth at young age can affect pre-college GPA measures as more learning resources are available. Their effect on net worth later in life is undeniable. Apart from net worth, other family background characteristics likely affect both pre-college test scores and net worth later in life. We include parental education, maternal age at first birth and the individual's birth, as well as the number of siblings to capture family background characteristics. Individual-specific characteristics are other important confounders. We include sex and citizenship status based on birth as further covariates. Conditional on this rich set of covariates X, and the unobserved variable ability U, there is no reason to believe that pre-college test

scores Z would affect or be correlated with post-college earnings through any other channel than obtaining a BA degree A. Despite our best efforts in explaining U, and the provided arguments in favour of assumption 8.1.2a, a test or conditional instrument exogeneity is not possible. A specification test is not feasible, because in this example the model is not overidentified.

8.5   Estimation

Estimation of the fully linear model is now straightforward. As in Tien [2022], we call the estimator an *instrumented common confounding* (ICC) estimator.

$$\hat{\beta}_{ICC} = \left(A^\intercal P_Z M_{T,X} A\right)^{-1} \left(A^\intercal P_Z M_{T,X} Y\right). \tag{8.18}$$

Here, $P_Z = Z\left(Z^\intercal Z\right)^{-1} Z^\intercal$ is the projection matrix of Z, and $M_{T,X} = I_n - P_{T,X}$ is the annihilator matrix of (T, X). In table 3, the estimates of four major methods are compared: ordinary least squares (OLS), instrumental variables (IV), proximal learning (PL), and the here suggested ICC estimator. The row corresponding to T describes the estimated partial effect of T (normalised to standard deviation one) on net worth Y (at 35) in the respective regressions. T is only used in proximal learning and ICC, but derived from the covariation of (Z, A) and W in negative control [Cui et al., 2020], as opposed to Z and W in our approach. The row corresponding to A contains estimates for β, the causal effect of obtaining a BA degree A on net worth Y (at 35). Their unit is US Dollar.

Table 3: Estimates with different estimators (Y in thousands (k))

	OLS	PL	IV	ICC
A	59.18***	30.90***	222.97***	125.15**
	(9.12)	(10.40)	(34.74)	(52.93)
T		27.76***		16.05**
		(4.82)		(7.37)

Notes: The table contains estimates and their standard errors (in parantheses) for β in the A row, and the linear parameter on T if used in the method from four estimators: Ordinary Least Squares (OLS), Proximal Learning (NC), Instrumental Variables (IV), and Instrumented Common Confounding (ICC). Asterisks indicate significance at the 1% (***), 5% (**) and 10% (*) level.

OLS estimates that obtaining a BA degree increases net worth at 35 by 59k\$. The

proximal learning estimator conditions on its own T, so implicitly anything fixed that covaries (Z, A) and proxies W. The proximal learning estimate at 31k\$, is indeed economically significantly smaller than the OLS estimate. As hypothesised, this might indicate that unobserved ability, which correlates (Z, A) and W, is a confounder which biases the estimated effect of education on net worth upwards. In contrast, the IV estimate is much larger at 223k\$. The inherent negative selection bias may thus be quite large. However, the IV estimator ignores the strong correlation of the pre-college test score instruments Z with ability U, which may lead to an accentuated ability bias compared to that in OLS. As the estimator should be robust to ability bias, we condition on T and obtain the ICC estimator at 125k\$. Indeed, conditioning on T attenuates the estimate by the expected ability bias. As relevance is not strongly satisfied for the instruments Z conditional on T in the ICC estimator, the standard error is expectedly large for this method. Still, both general selection bias and ability bias appear to be strong confounders in this difficult identification problem.

Quantitatively separating ability and general selection bias helps add the necessary credibility to IV, which misses under the original IV exogeneity assumption.

9   Conclusion

In this work, we relax instrument exogeneity in the presence of mismeasured confounders. Other observed variables, the proxies, must be relevant for the unobserved confounders, which cause endogeneity in the instruments. The mild parametric index sufficiency assumption is also required. Importantly, the proxies can be economically meaningful variables, with their own effects on treatment and outcome. This method can be useful in various causal identification problems with observational data, where the unobserved confounders are otherwise unrestricted observed variables. The linear returns to education identification problem illustrates how this method can identify causal effects when instrument exogeneity, as often in practice, is a strong and hardly testable assumption.

This paper established two point identification results. When point identification is impossible, this approach can still identify informative bounds on causal effects. This set identification exercise is left to future work. Further, we have not demonstrated how to construct estimators other than in the linear example. Uncertainty in the control function estimation will be reflected in the performance of any estimator using this identification approach. The integration of this approach, which at best identifies averages of causal effects across unobservables, with marginal treatment effects, is another remaining task.

References

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. *arXiv e-prints*, pages arXiv–2208, 2022.

Richard W Blundell and James L Powell. Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679, 2004.

U.S. Department of Labor Bureau of Labor Statistics. National longitudinal survey of youth 1997 cohort, 1997-2017 (rounds 1-18), 2019.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Chapter 77 linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*, 2020.

Xavier D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.

Xavier D'Haultfœuille, Stefan Hoderlein, and Yuya Sasaki. Testing and relaxing the exclusion restriction in the control function approach. *Journal of Econometrics*, 2021.

Zvi Griliches. Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, pages 1–22, 1977.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

James J Heckman, Lance J Lochner, and Petra E Todd. Earnings functions, rates of return and treatment effects: The mincer equation and beyond. *Handbook of the Economics of Education*, 1:307–458, 2006.

Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

Roger Klein and Francis Vella. Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics*, 154(2):154–164, 2010.

Arthur Lewbel. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1):67–80, 2012.

Laura Liu, Alexandre Poirier, and Ji-Liang Shiu. Identification and estimation of average partial effects in semiparametric binary response panel models. *arXiv preprint arXiv:2105.12891*, 2021.

Rolf Loeber, Barbara Menting, Donald R Lynam, Terri E Moffitt, Magda Stouthamer-Loeber, Rebecca Stallings, David P Farrington, and Dustin Pardini. Findings from the pittsburgh youth study: Cognitive impulsivity and intelligence as predictors of the age–crime curve. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(11):1136–1149, 2012.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Daniel L Millimet and Rusty Tchernis. Estimation of treatment effects without an exclusion restriction: With an application to the analysis of the school breakfast program. *Journal of Applied Econometrics*, 28(6):982–1017, 2013.

Kenichi Nagasawa. Treatment effect estimation with noisy conditioning variables. *arXiv preprint arXiv:1811.00667*, 2018.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Whitney K Newey, James L Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

George Psacharopoulos and Harry Anthony Patrinos. Returns to investment in education: a decennial review of the global literature. *Education Economics*, 26(5):445–458, 2018.

Christian Tien. Instrumented common confounding. 2022.

Emmanuel Selorm Tsyawo. Feasible iv regression without excluded instruments. *arXiv preprint arXiv:2103.09621*, 2021.

10 Proofs

Section 3 proofs

Proof of lemma 1. Let $t \in \mathcal{Z}^\tau$.

$$
\begin{aligned}
f_{W,Z|T}(W, Z|T) &= \int_{\mathcal{U}} f_{W,Z|U,T}(W, Z|u, T) f_{U|T}(u, T) \, \mathrm{d}\mu_U(u) \\
&= \int_{\mathcal{U}} f_{W|Z,U,T}(W|Z, u, T) f_{Z|U,T}(Z|u, T) f_{U|T}(u, T) \, \mathrm{d}\mu_U(u) \\
&= \int_{\mathcal{U}} f_{W|U}(W|u) f_{Z|T}(Z|T) f_{U|T}(u, T) \, \mathrm{d}\mu_U(u) \\
&= f_{Z|T}(Z|T) \int_{\mathcal{U}} f_{W|U}(W|u) f_{U|T}(u, T) \, \mathrm{d}\mu_U(u) \\
&= f_{Z|T}(Z|T) f_{W|T}(W|T)
\end{aligned}
$$

From line two to three we use $W \perp\!\!\!\perp Z \mid U$ (assumption 2.1.3a) and $U \perp\!\!\!\perp Z \mid T$ (by construction of T). From line four to five we again use $W \perp\!\!\!\perp \tau(Z) \mid U$ (assumption 2.1.3a). $\qquad \square$

Proof of lemma 2. We write $f_{W|Z}(W, Z)$ in two separate ways using T and relate those.

1.

$$
\begin{aligned}
f_{W|Z}(W, Z) &= \int_{\mathcal{U}} f_{W|U}(W, u) f_{U|Z}(u, Z) \, \mathrm{d}\mu_U(u) \\
&= \int_{\mathcal{U}} f_{W|U}(W|u) f_{U|T,Z}(u, t, z) \, \mathrm{d}\mu_U(u)
\end{aligned}
$$

The first line follows from $W \perp\!\!\!\perp Z \mid U$ (assumption 2.1.3a).

2.

$$
\begin{aligned}
f_{W|Z}(W, Z) &= f_{W|T}(W, T) \\
&= \int_{\mathcal{U}} f_{W|U}(W, u) f_{U|T}(u, t) \, \mathrm{d}\mu_U(u)
\end{aligned}
$$

The first line follows from the construction of $T = \tau(Z)$, $\tau \in \mathcal{L}_2(Z)$, such that $W \perp\!\!\!\perp Z \mid T$.

By the equality of the expressions in steps 1 and 2 above, for all $z \in \mathcal{Z}$,

$$
\int_{\mathcal{U}} f_{W|U}(W, u) f_{U|T,Z}(u, t, z) \, \mathrm{d}\mu_U(u) = \int_{\mathcal{U}} f_{W|U}(W, u) f_{U|T}(u, t) \, \mathrm{d}\mu_U(u),
$$

$$
\int_{\mathcal{U}} f_{U|T,Z}(u, t, z) \frac{f_W(W)}{f_U(u)} f_{U|W}(u, W) \, \mathrm{d}\mu_U(u) = \int_{\mathcal{U}} f_{U|T}(u, t) \frac{f_W(W)}{f_U(u)} f_{U|W}(u, W) \, \mathrm{d}\mu_U(u).
$$

44

Let $g_{t,z}(U) := \frac{f_{U|T,Z}(U,t,z)}{f_U(U)}$ and $g_t(U) := \frac{f_{U|T}(U,t)}{f_U(U)}$ for any $z \in \mathcal{Z}$. Then,

$$\mathbb{E}\left[g_{t,z}(U)f_W(W)|W\right] = \mathbb{E}\left[g_t(U)f_W(W)|W\right]$$

$$\mathbb{E}\left[\left(g_{t,z}(U) - g_t(U)\right)\Big| W\right] f_W(W) = 0$$

By completeness of W for U (assumption 2.1.3b), the above only holds if

$$g_{t,z}(U) = g_t(U).$$

This implies

$$f_{U|T,Z}(U,t,z) = f_{U|T}(U,t),$$

and thus $U \perp\!\!\!\perp Z \mid T$ for any $T = \tau(Z)$, $\tau \in \mathcal{L}_2(Z)$ such that $W \perp\!\!\!\perp Z \mid T$. $\qquad\square$

Section 4.1 proofs

Proof of Theorem 1.

$$\begin{aligned}
\mathbb{E}\left[Y|Z\right] &= \mathbb{E}\left[k_0(A) + \varepsilon \mid Z\right] \\
&= \mathbb{E}\left[k(A)|Z\right] + \mathbb{E}\left[\mathbb{E}\left[\varepsilon|U\right]|Z\right] \\
&= \mathbb{E}\left[k(A)|Z\right] + \mathbb{E}\left[\mathbb{E}\left[\varepsilon|U\right]|T\right] \\
&= \mathbb{E}\left[k(A)|Z\right] + \mathbb{E}\left[\varepsilon|T\right] \\
&= \mathbb{E}\left[k(A) + \mathbb{E}\left[\varepsilon|T\right]|Z\right] \\
h(A,T) &= k(A) + \mathbb{E}\left[\varepsilon|T\right]
\end{aligned}$$

From line one to two we use the moment $\mathbb{E}\left[\varepsilon|Z,U\right] = \mathbb{E}\left[\varepsilon|U\right]$ formulated in assumption 4.1. From line two to three we use $U \perp\!\!\!\perp Z \mid T$ and $T = \tau(Z)$ (assumption 2.1.2b/2c). Required for this step is the identifiability of T, given by lemma 1 subject to assumption 2.1.3. Completeness of Z for A given T (assumption 2.1.2c) is used to obtain the final line

given that $\mathbb{E}[Y|Z] = \mathbb{E}[h(A, T)|Z]$. It follows that

$$
\begin{aligned}
\int_{\mathcal{A}} Y(a)\pi(a)\mathrm{d}a &= \int_{\mathcal{A}} k(a)\pi(a)\mathrm{d}a \\
&= \int_{\mathcal{A}} \left(h(a, t) - \mathbb{E}[\varepsilon|t] \right) \pi(a)\mathrm{d}a \text{ for any } t \in \mathcal{T} \\
&= \int_{\mathcal{T}} \int_{\mathcal{A}} \left(h(a, t) - \mathbb{E}[\varepsilon|t] \right) \pi(a)\mathrm{d}a f(t)\mathrm{d}t \\
&= \int_{\mathcal{T}} \int_{\mathcal{A}} h(a, t)\pi(a)\mathrm{d}a f(t)\mathrm{d}t.
\end{aligned}
$$

In line one we use $\int_{\varepsilon} \varepsilon f(\varepsilon)\mathrm{d}\varepsilon = 0$. From line one to two we use $h(A, T) = k(A) + \mathbb{E}[\varepsilon|T]$. From line two to three we simply extend what holds for any given t to all $t \in \mathcal{T}$. From line three to four we use $\int_{\mathcal{T}} \mathbb{E}[\varepsilon|t] f(t)\mathrm{d}t = \mathbb{E}[\varepsilon] = 0$.

$\square$

Section 4.2 proofs

Proof of lemma 3.

$$
F_{\eta|T}(\eta) = \int_{\mathcal{U}} F_{\eta|u}(\eta) f_{U|T}(u, T)\,\mathrm{d}\mu_U(u),
$$

$$
\left.\frac{\partial F_{\eta|T}(e)}{\partial e}\right|_{e=\eta} = \int_{\mathcal{U}} \underbrace{\left.\frac{\partial F_{\eta|u}(e)}{\partial e}\right|_{e=\eta}}_{>0\ \forall u, \eta} f_{U|T}(u, T)\,\mathrm{d}\mu_U(u) > 0,
$$

so $F_{\eta|T}(\eta)$ is strictly increasing on the conditional support of η.

Now prove $Z \perp\!\!\!\perp \eta \mid T$.

$$
\begin{aligned}
f_{Z,\eta|T} &= \int_{\mathcal{U}} f_{Z,\eta|U,T} f_{U|T}\,\mathrm{d}\mu_U(u) \\
&= \int_{\mathcal{U}} f_{\eta|Z,U,T} f_{Z|U,T} f_{U|T}\,\mathrm{d}\mu_U(u) \\
&= \int_{\mathcal{U}} f_{\eta|U,T} f_{Z|T} f_{U|T}\,\mathrm{d}\mu_U(u) \\
&= f_{Z|T} \int_{\mathcal{U}} f_{\eta|U,T} f_{U|T}\,\mathrm{d}\mu_U(u) \\
&= f_{Z|T} f_{\eta|T}
\end{aligned}
$$

From second to third line we use $Z \perp\!\!\!\perp \eta \mid U$ (assumption 4.2.3) and $Z \perp\!\!\!\perp U \mid T$ (assumption 2.1.4). We have shown that $Z \perp\!\!\!\perp \eta \mid T$.

$\square$

Proof of theorem 2. First, we derive a preliminary result as if U were observed.

$$
\begin{aligned}
F_{A|Z,U}(a, z, u) = \Pr\left(A \le a | Z = z, U = u\right) &= \Pr\left(h(z, \eta) \le a | Z = z, U = u\right) \\
&= \Pr\left(\eta \le h^{-1}(a, z) | Z = z, U = u\right) \\
&= \Pr\left(\eta \le h^{-1}(a, z) | U = u\right) \\
&= F_{\eta|u}\left(h^{-1}(a, z)\right) \\
&= F_{\eta|u}\left(\eta\right).
\end{aligned}
$$

From line one to two we use the invertibility of $h(z, \eta)$ (assumption 4.2.(1/2)). From line two to three we use $Z \perp\!\!\!\perp \eta \mid U$ (assumption 4.2.3). Then,

$$
\begin{aligned}
V_T := F_{A|Z}(A, Z) = \int_{\mathcal{U}} F_{A|Z,U}(A, Z, u) f_{U|T}(u, T) \, \mathrm{d}\mu_U(u) \\
= \int_{\mathcal{U}} F_{\eta|U}\left(\eta\right) f_{U|T}(u, T) \, \mathrm{d}\mu_U(u) \\
= F_{\eta|T}(\eta)
\end{aligned}
$$

On line one we use $Z \perp\!\!\!\perp U \mid T$ (assumption 2.1.2c). From line one to two, we use the result derived at the beginning of this proof. The final line again follows from $\tau(Z) \perp\!\!\!\perp \eta \mid U$ (assumption 4.2.3). Hence,

$$
V_T = F_{A|Z}(A, Z) = F_{\eta|T}(\eta).
$$

By lemma 3 (strictly increasing $F_{\eta|T}$ on the support of η), (η, T) and $(V_T, T) = (F_{\eta|T}(\eta), T)$ are associated with the same sigma algebra.

We show $A \perp\!\!\!\perp Y(a) \mid (V_T, T)$.

$$
\begin{aligned}
&f_{Y(a),A|V_T,T}(Y(a), A | V_T, T) \\
&= \int_{\mathcal{U}} \underbrace{f_{Y(a)|A,V_T,T,U}(Y(a)|h(Z, \eta), V_T, T, u)}_{=f_{Y(a)|V_T,T,u}(Y(a)|V_T,T,u)} \underbrace{f_{A|T,V_T,U}(h(Z, \eta)|V_T, T, u)}_{=f_{A|V_T,T}(h(Z,\eta)|V_T,T)} f_{U|T}(u|T) \, \mathrm{d}\mu_U(u) \\
&= f_{A|V_T,T}(h(Z, \eta)|V_T, T) \underbrace{\int_{\mathcal{U}} f_{Y(a)|V_T,T,u}(Y(a)|V_T, T, u) f_{U|T}(u|T) \, \mathrm{d}\mu_U(u)}_{=f_{Y(a)|V_T,T}(Y(a)|V_T,T)} \\
&= f_{A|V_T,T}(A|V_T, T) f_{Y(a)|V_T,T}(Y(a)|V_T, T) \implies A \perp\!\!\!\perp Y(a) \mid (V_T, T)
\end{aligned}
$$

$\square$

Proof of theorem 3. $\theta_0 := \mathbb{E}\left[\int_{\mathcal{A}} Y(a)\pi(a)\,\mathrm{d}\mu_A(a)\right]$ is identified as

$$\int_{\mathcal{V}_T, \mathcal{Z}^\tau} \int_{\mathcal{A}} \mathbb{E}\left[Y|A = a, (V_T, T) = (v_T, t)\right] \pi(a)\,\mathrm{d}\mu_A(a)\,\mathrm{d}F_{V_T,T}(v_T, t)$$

$$\mathbb{E}\left[\int_{\mathcal{A}} \mathbb{E}\left[Y(a)|A = a, V_T, T\right]\pi(a)\,\mathrm{d}\mu_A(a)\right]$$

$$\mathbb{E}\left[\int_{\mathcal{A}} \mathbb{E}\left[g(A,\varepsilon)|A = a, V_T, T\right]\pi(a)\,\mathrm{d}\mu_A(a)\right]$$

$$\mathbb{E}\left[\int_{\mathcal{A}} \int_{\mathcal{E}} g(A,\varepsilon)\,\mathrm{d}F_{\varepsilon|A,V_T,T}(\varepsilon, a, V_T, T)\pi(a)\,\mathrm{d}\mu_A(a)\right]$$

$$\mathbb{E}\left[\int_{\mathcal{A}} \int_{\mathcal{E}} g(A,\varepsilon)\,\mathrm{d}F_{\varepsilon|V_T,T}(\varepsilon, V_T, T)\pi(a)\,\mathrm{d}\mu_A(a)\right]$$

$$\mathbb{E}\left[\int_{\mathcal{E}} \int_{\mathcal{A}} g(A,\varepsilon)\pi(a)\,\mathrm{d}\mu_A(a)\,\mathrm{d}F_{\varepsilon|V_T,T}(\varepsilon, V_T, T)\right]$$

$$\mathbb{E}\left[\int_{\mathcal{A}} g(A,\varepsilon)\pi(a)\,\mathrm{d}\mu_A(a)\right]$$

$$\mathbb{E}\left[\int_{\mathcal{A}} Y(a)\pi(a)\,\mathrm{d}\mu_A(a)\right] = \theta_0.$$

We let $Y(a) = g(a,\varepsilon)$ for some $\varepsilon \in \mathcal{E}$ and $g \in \mathcal{L}_2(A, \varepsilon)$ from line two to three. Then, $A \perp\!\!\!\perp Y(a) \mid (V_T, T)$ from theorem 2 implies $A \perp\!\!\!\perp \varepsilon \mid (V_T, T)$, which we use from line four to five. All other steps are algebra. $\qquad\square$

Section 6 proofs

Proof of theorem 4. For any $\tau_0 \in \mathcal{T}_{\text{valid}}$,

$$\mathbb{E}\left[m(O; \tau_0 + \eta) - m(O; \tau_0)\right] = \mathbb{E}\left[m(O; \eta)\right]$$
$$= \mathbb{E}\left[\alpha(Z)\eta(Z)\right].$$

The first line holds by linearity of $\tau \mapsto \mathbb{E}\left[m(O; \tau)\right]$, the second by the Riesz representation theorem.

For any $\tau_0 \in \mathcal{T}_{\text{valid}}$, consider any $\eta \in \mathcal{N}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}})$, i.e.

$$\Pi_{\mathcal{T}}\left[\mathbb{E}\left[\tau_0(Z) + \eta(Z)|U\right]|Z\right] = 0.$$

Completeness of Z for U by construction implies $\mathbb{E}\left[h(\eta(Z); g)|U\right] = 0$.

For any $\eta \in \mathcal{N}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}})$ instruments Z remain exogenous, which implies

$$0 = \mathbb{E}\left[m(O; \eta)\right] = \mathbb{E}\left[\alpha(Z)\eta(Z)\right].$$

The above condition holds for any $\eta \in \mathcal{N}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}})$, therefore it must be that $\alpha \in \mathcal{N}^{\perp}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}})$ when θ_0 is identified as $\theta_0 = \mathbb{E}\left[m(O; \tau_0)\right]$ for any $\tau_0 \in \mathcal{T}_{\text{valid}}$.

Note that by $\mathbb{E}\left[g(Z)|W\right] = \mathbb{E}\left[\mathbb{E}\left[g(Z)|U\right]|W\right]$ for any $g \in \mathcal{G}$ and $\Pi_{\mathcal{T}}\left[q(W)|Z\right] = \Pi_{\mathcal{T}}\left[\mathbb{E}\left[q(W)|U\right]|Z\right]$ for $\mathcal{T} \subseteq \mathcal{L}_2(Z)$ for any $q \in \mathcal{L}_2(W)$ (relaxed version of assumption 2.1.3a),

$$
\begin{aligned}
P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)} P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}} &= \Pi_{\mathcal{T}}\left[\mathbb{E}\left[\tau_0(Z) + \eta(Z)|W\right]|Z\right] \\
&= \Pi_{\mathcal{T}}\left[\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\tau_0(Z) + \eta(Z)|U\right]|W\right]|U\right]|Z\right] = P_{Z,\mathcal{T}}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{\mathcal{L}_2(W)} P_{\mathcal{L}_2(W)}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}}.
\end{aligned}
$$

Then, by completeness of W for U (assumption 2.1.3b),

$$
\begin{aligned}
P_{\mathcal{L}_2(U)}^{\mathcal{L}_2(W)} P_{\mathcal{L}_2(W)}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}} &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\tau_0(Z) + \eta(Z)|U\right]|W\right]|U\right] \\
&= \mathbb{E}\left[\tau_0(Z) + \eta(Z)|U\right] = P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}}.
\end{aligned}
$$

Completeness of W for U (assumption 2.1.3b) implies the above for the following reason: Only if completeness is violated can it be that $0 = \mathbb{E}\left[g(U)|W\right]$ (and thus $0 = \mathbb{E}\left[\mathbb{E}\left[g(U)|W\right]|U\right]$) for some $g(U) \neq 0$. Consequently, $\gamma(U) = \mathbb{E}\left[\mathbb{E}\left[\gamma(U) + g(U)|W\right]|U\right]$ only if $g(U) = 0$ when completeness holds, which implies $\gamma(U) = \mathbb{E}\left[\mathbb{E}\left[\gamma(U)|W\right]|U\right]$ when completeness holds.

Jointly, this concludes the proof as

$$
\alpha \in \mathcal{N}^{\perp}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(U)} P_{\mathcal{L}_2(U)}^{Z,\mathcal{T}}) = \mathcal{N}^{\perp}(P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)} P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}).
$$

$\square$

Proof of theorem 5. θ_0 is strongly identified, so

$$
\mathbb{E}\left[q_0(W)\tau(Z)\right] = \mathbb{E}\left[\alpha(Z)\tau(Z)\right] = \mathbb{E}\left[m(O; \tau)\right] \text{ for any } h \in \mathcal{H}.
$$

Use this, and $\mathbb{E}\left[g_0(Z)|W\right] - \mathbb{E}\left[\tau_0(Z)|W\right]$ below:

$$
\begin{aligned}
\mathbb{E}\left[\psi(O;\tau,q)\right] - \theta_0 &= \mathbb{E}\left[m(O;\tau) + q(W)(g_0(Z) - \tau(Z))\right] - \theta_0 \\
&= \mathbb{E}\left[m(O;\tau) + q(W)(g_0(Z) - \tau(Z))\right] - \mathbb{E}\left[m(O;\tau_0)\right] - \mathbb{E}\left[q(W)(g_0(Z) - \tau_0(Z))\right] \\
&= \mathbb{E}\left[\alpha(Z)\tau(Z)\right] - \mathbb{E}\left[\alpha(Z)\tau_0(Z)\right] + \mathbb{E}\left[q(W)(\tau_0(Z) - \tau(Z))\right] \\
&= \mathbb{E}\left[\alpha(Z)(\tau(Z) - \tau_0(Z))\right] - \mathbb{E}\left[q(W)(\tau(Z) - \tau_0(Z))\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[q_0(W)|Z\right](\tau(Z) - \tau_0(Z))\right] - \mathbb{E}\left[q(W)(\tau(Z) - \tau_0(Z))\right] \\
&= \mathbb{E}\left[q_0(W)(\tau(Z) - \tau_0(Z))\right] - \mathbb{E}\left[q(W)(\tau(Z) - \tau_0(Z))\right] \\
&= -\mathbb{E}\left[(q(W) - q_0(W))(\tau(Z) - \tau_0(Z))\right]
\end{aligned}
$$

By projecting onto Z and W respectively, the error rates follow:

$$
\begin{aligned}
\left|\mathbb{E}\left[\psi(O;\tau,q)\right] - \theta_0\right| &= \left|\mathbb{E}\left[\mathbb{E}\left[q(W) - q_0(W)|Z\right](\tau(Z) - \tau_0(Z))\right]\right| = \left|\left\langle P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}(q - q_0), \tau - \tau_0\right\rangle\right| \\
&= \left|\mathbb{E}\left[(q(W) - q_0(W))\mathbb{E}\left[\tau(Z) - \tau_0(Z)|W\right]\right]\right| = \left|\left\langle q - q_0, P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}(\tau - \tau_0)\right\rangle\right|
\end{aligned}
$$

Then use the Cauchy-Schwartz inequality to obtain

$$
\left|\mathbb{E}\left[\psi(O;\tau,q)\right] - \theta_0\right| \leq \min\left\{\left\|P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}(\tau - \tau_0)\right\|_2 \|q - q_0\|_2, \ \|\tau - \tau_0\|_2 \left\|P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}(q - q_0)\right\|_2\right\}
$$

Double robustness is satisfied as

$$
\begin{aligned}
\mathbb{E}\left[\psi(O;\tau,q_0)\right] &= \theta_0 - \mathbb{E}\left[(q_0(W) - q_0(W))(\tau(Z) - \tau_0(Z))\right] = \theta_0 \\
\mathbb{E}\left[\psi(O;\tau_0,q)\right] &= \theta_0 - \mathbb{E}\left[(q(W) - q_0(W))(\tau_0(Z) - \tau_0(Z))\right] = \theta_0.
\end{aligned}
$$

Neyman orthogonality is satisfied as $\tau_0 \in \mathcal{T}_{\text{valid}}$ and $q_0 \in \mathcal{Q}_0$ if and only if

$$
\begin{aligned}
\frac{\partial}{\partial t}\mathbb{E}\left[\psi(O;\tau_0 + t\tau, q_0)\right]\Big|_{t=0} &= \mathbb{E}\left[(\alpha(Z) - q_0(W))\tau(Z)\right] = \mathbb{E}\left[\mathbb{E}\left[\alpha(Z) - q_0(W)|Z\right]\tau(Z)\right] = 0 \\
\frac{\partial}{\partial t}\mathbb{E}\left[\psi(O;\tau, q_0 + tq)\right]\Big|_{t=0} &= \mathbb{E}\left[q(W)(g_0(Z) - \tau_0(Z))\right] = \mathbb{E}\left[q(W)\mathbb{E}\left[g_0(Z) - \tau_0(Z)|W\right]\right] = 0.
\end{aligned}
$$

$\square$

Proof of theorem 6. Want to show that:

$$\mathbb{E}\left[m_3(O;k,\tau,g,q_k,q_\tau,\alpha_g)\right] - \theta_0$$

$$= \mathbb{E}\left[m_0(O;k) + q_k(Z)\left(g(Z) - \tau(Z) - k(A;\tau,g)\right) + q_\tau(W;q_k)\left(\tau(Z;g) - g(Z)\right) + \alpha_g(Z;q_k,q_\tau)\left(g_Y(Y) - g(Z)\right)\right] - \theta_0$$

$$= \mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\left(k(A;\tau,g) - k_0(A;\tau_0,g_0)\right)\right]$$

$$+ \mathbb{E}\left[\left(q_{\tau,0}(W;q_k) - q_\tau(W;q_k)\right)\left(\tau_0(Z;g_0) - \tau(Z;g)\right)\right]$$

$$+ \mathbb{E}\left[\left(\alpha_{g,0}(Z;q_k,q_\tau) - \alpha_g(Z;q_k,q_\tau)\right)\left(g(Z) - g_0(Z)\right)\right]$$

First, deal with $\mathbb{E}\left[m_0(O;k) + q_k(Z)\left(g(Z) - \tau(Z) - k(A)\right)\right] - \theta_0$. From line one to two below use that $\theta_0 = \mathbb{E}\left[m(O;k_0)\right]$, and that $\mathbb{E}\left[m(O;k)\right] = \mathbb{E}\left[\alpha_{k,0}k(A)\right]$ for any $k \in \mathcal{K}$ by the Riesz representation theorem when $k \mapsto \mathbb{E}\left[m(O;k)\right]$ is a linear and continuous functional. From line one to two also use that $\mathbb{E}\left[k_0(A;\tau,g)|Z\right] = g(Z) - \tau(Z)$ for any $\tau \in \mathcal{T}$ and $g \in \mathcal{G}$. From line two to three use that $\alpha_{k,0}(A) = \mathbb{E}\left[q_{k,0}(Z)|A\right]$ for $q_{k,0}(Z) \in \mathcal{Q}_{k,0}$ under strong instrument relevance 6.1.

$$\mathbb{E}\left[m_0(O;k(\tau,g)) + q_k(Z)\left(g(Z) - \tau(Z) - k(A;\tau,g)\right)\right] - \theta_0$$

$$= \mathbb{E}\left[\alpha_{k,0}(A)k(A;\tau,g)\right] - \mathbb{E}\left[\alpha_{k,0}(A)k_0(A;\tau_0,g_0)\right]$$

$$+ \mathbb{E}\left[q_k(Z)\left(g(Z) - \tau(Z;g) - k(A;\tau,g)\right)\right] - \mathbb{E}\left[q_k(Z)\left(g_0(Z) - \tau_0(Z;g_0) - k_0(A;\tau_0,g_0)\right)\right]$$

$$= \mathbb{E}\left[q_{k,0}(Z)\left(k(A;\tau,g) - k_0(A;\tau_0,g_0)\right)\right]$$

$$+ \mathbb{E}\left[q_k(Z)\left(\left(g(Z) - \tau(Z;g) - k(A;\tau,g)\right) - \left(g_0(Z) - \tau_0(Z;g_0) - k_0(A;\tau_0,g_0)\right)\right)\right]$$

$$= \mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\left(k(A;\tau,g) - k_0(A;\tau_0,g_0)\right)\right]$$

$$+ \mathbb{E}\left[q_k(Z)\left(\left(g(Z) - \tau(Z;g)\right) - \left(g_0(Z) - \tau_0(Z;g_0)\right)\right)\right]$$

Now, deal with the terms involving τ_0 and τ from the previous step, and $\mathbb{E}\left[q_\tau(W;q_k)\left(\tau(Z;g) - g(Z)\right)\right]$ from m_3. From line one to two use theorem 4, which implies $\mathbb{E}\left[q_k(Z)\tau(Z;g)\right] = \mathbb{E}\left[\alpha_{\tau,0}(Z;q_k)\tau(Z;g)\right]$ with $\alpha_{\tau,0}(q_k) \in \mathcal{N}^\perp(P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}})$. Also, from line one to two use $\mathbb{E}\left[g(Z)|W\right] = \mathbb{E}\left[\tau_0(Z;g)|W\right]$ for any $g \in \mathcal{G}$.

$$\mathbb{E}\left[q_k(Z)\left(\tau_0(Z;g_0) - \tau(Z;g)\right)\right] + \mathbb{E}\left[q_\tau(W;q_k)\left(\tau(Z;g) - g(Z)\right)\right]$$

$$= \mathbb{E}\left[q_{\tau,0}(W;q_k)\left(\tau_0(Z;g_0) - \tau(Z;g)\right)\right] + \mathbb{E}\left[q_\tau(W;q_k)\left(\tau(Z;g) - g(Z)\right)\right] - \mathbb{E}\left[q_\tau(W;q_k)\left(\tau_0(Z;g_0) - g_0(Z)\right)\right]$$

$$= \mathbb{E}\left[\left(q_{\tau,0}(W;q_k) - q_\tau(W;q_k)\right)\left(\tau_0(Z;g_0) - \tau(Z;g)\right)\right] + \mathbb{E}\left[q_\tau(W;q_k)\left(g_0(Z) - g(Z)\right)\right]$$

Finally, deal with the remaining terms involving g and g_0. From line to two use that

$\alpha_{g,0}(Z; q_k, q_\tau) = \mathbb{E}\left[(q_k(Z) - q_\tau(W; q_k)) \,|Z\right]$. Also use that $\mathbb{E}\left[g_Y(Y)|Z\right] = g_0(Z)$.

$$\mathbb{E}\left[(q_k(Z) - q_\tau(W; q_k))(g(Z) - g_0(Z))\right] + \mathbb{E}\left[\alpha_g(Z; q_k, q_\tau)(g_Y(Y) - g(Z))\right]$$
$$= \mathbb{E}\left[\alpha_{g,0}(Z; q_k, q_\tau)(g(Z) - g_0(Z))\right] + \mathbb{E}\left[\alpha_g(Z; q_k, q_\tau)(g_Y(Y) - g(Z))\right] - \mathbb{E}\left[\alpha_g(Z; q_k, q_\tau)(g_Y(Y) - g_0(Z))\right]$$
$$= \mathbb{E}\left[\alpha_{g,0}(Z; q_k, q_\tau)(g(Z) - g_0(Z))\right] + \mathbb{E}\left[\alpha_g(Z; q_k, q_\tau)(g_0(Z) - g(Z))\right]$$
$$= \mathbb{E}\left[(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau))(g(Z) - g_0(Z))\right]$$

Join all of the above to complete the proof:

$$\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g)\right] - \theta_0 = \mathbb{E}\left[(q_{k,0}(Z) - q_k(Z))(k(A) - k_0(A))\right]$$
$$+ \mathbb{E}\left[(q_{\tau,0}(W; q_k) - q_\tau(W; q_k))(\tau_0(Z) - \tau(Z))\right]$$
$$+ \mathbb{E}\left[(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau))(g(Z) - g_0(Z))\right].$$

$\square$

Proof of corollary 1. The conditions of theorem hold, 6 hold, so

$$\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g)\right] - \theta_0 = \underbrace{\mathbb{E}\left[(q_{k,0}(Z) - q_k(Z))(k(A; \tau, g) - k_0(A; \tau_0, g_0))\right]}_{\text{Term 1}}$$
$$+ \underbrace{\mathbb{E}\left[(q_{\tau,0}(W; q_k) - q_\tau(W; q_k))(\tau_0(Z; g_0) - \tau(Z; g))\right]}_{\text{Term 2}}$$
$$+ \underbrace{\mathbb{E}\left[(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau))(g(Z) - g_0(Z))\right]}_{\text{Term 3}}.$$

1. Term 1 is zero whenever $\left((k = k_0(\tau_0, g_0)) \vee (q_k = q_{k,0})\right)$, irrespective of the other two terms.

2. Term 2 is zero whenever $\left((\tau = \tau_0(g_0)) \vee (q_\tau = q_{\tau,0})\right)$, irrespective of the other two terms.

3. Term 3 is zero whenever $\left((g = g_0) \vee (\alpha_g = \alpha_{g,0})\right)$, irrespective of the other two terms.

This completes the proof. $\square$

Proof of corollary 2. The conditions of theorem hold, 6 hold, so

$$\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g)\right] - \theta_0 = \mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\left(k(A; \tau, g) - k_0(A; \tau_0, g_0)\right)\right]$$
$$+ \mathbb{E}\left[\left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right)\left(\tau_0(Z; g_0) - \tau(Z; g)\right)\right]$$
$$+ \mathbb{E}\left[\left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right)\left(g(Z) - g_0(Z)\right)\right].$$

Let $\partial_g f(g, h) := \left.\frac{\partial}{\partial t} f(.; g_0 + tg, h)\right|_{t=0}$ for any functions f, g, and h.

$$\left.\frac{\partial}{\partial t}\mathbb{E}\left[m_3(O; k_0 + tk, \tau, g, q_k, q_\tau, \alpha_g)\right]\right|_{t=0} = \mathbb{E}\left[q_{k,0}(Z) - q_k(Z)\right] = 0 \text{ if } q_k = q_{k,0},$$

$$\left.\frac{\partial}{\partial t}\mathbb{E}\left[m_3(O; k, \tau_0 + t\tau, g, q_k, q_\tau, \alpha_g)\right]\right|_{t=0}$$
$$= \mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\partial_\tau k(A; \tau, g) + \left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right)\right]$$
$$= 0 \text{ if } (q_\tau = q_{\tau,0}) \wedge \left((\partial_\tau k(\tau, g) = 0) \vee \left(q_k = q_{k,0}\right)\right),$$

$$\left.\frac{\partial}{\partial t}\mathbb{E}\left[m_3(O; k, \tau, g_0 + tg, q_k, q_\tau, \alpha_g)\right]\right|_{t=0}$$
$$= \mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\partial_g k(A; \tau, g) - \left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right)\partial_g\tau(Z; g) + \left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right)\right]$$
$$= 0 \text{ if } (\alpha_g = \alpha_{g,0}) \wedge \left((\partial_g k(\tau, g) = 0) \vee (q_k = q_{k,0})\right) \wedge \left((\partial_g\tau(g) = 0) \vee (q_\tau = q_{\tau,0})\right),$$

$$\left.\frac{\partial}{\partial t}\mathbb{E}\left[m_3(O; k, \tau, g, q_{k,0} + tq_k, q_\tau, \alpha_g)\right]\right|_{t=0}$$
$$= \mathbb{E}\left[-\left(k(A; \tau, g) - k_0(A; \tau_0, g_0)\right) + \left(\partial_{q_k}\left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right)\right)\left(\tau_0(Z; g_0) - \tau(Z; g)\right)\right]$$
$$+ \mathbb{E}\left[\left(\partial_{q_k}\left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right)\right)\left(g(Z) - g_0(Z)\right)\right]$$
$$= 0 \text{ if } (k = k_0(\tau_0, g_0)) \wedge \left((\tau = \tau_0(g_0)) \vee (q_\tau = q_{\tau,0})\right) \wedge \left((g = g_0) \vee (\alpha_g = \alpha_{g,0})\right),$$

$$\left.\frac{\partial}{\partial t}\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_{\tau,0} + tq_\tau, \alpha_g)\right]\right|_{t=0}$$
$$= \mathbb{E}\left[-\left(\tau_0(Z; g_0) - \tau(Z; g)\right) + \partial_{q_\tau}\left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right)\left(g(Z) - g_0(Z)\right)\right]$$
$$= 0 \text{ if } (\tau = \tau_0(g_0)) \wedge \left((g = g_0) \vee (\alpha_g = \alpha_{g,0})\right),$$

$$\left.\frac{\partial}{\partial t}\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_{g,0} + t\alpha_g)\right]\right|_{t=0} = \mathbb{E}\left[-\left(g(Z) - g_0(Z)\right)\right] = 0 \text{ if } (g = g_0).$$

$\square$

Proof of corollary 3. The conditions of theorem hold, 6 hold, so

$$\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g)\right] - \theta_0 = \underbrace{\mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\left(k(A; \tau, g) - k_0(A; \tau_0, g_0)\right)\right]}_{\text{Term 1}}$$

$$+ \underbrace{\mathbb{E}\left[\left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right)\left(\tau_0(Z; g_0) - \tau(Z; g)\right)\right]}_{\text{Term 2}}$$

$$+ \underbrace{\mathbb{E}\left[\left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right)\left(g(Z) - g_0(Z)\right)\right]}_{\text{Term 3}}.$$

Each of these terms can be bounded using the Cauchy-Schwarz inequality.

Term 1:

$$\mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\left(k(A; \tau, g) - k_0(A; \tau_0, g_0)\right)\right]$$

$$= \mathbb{E}\left[\left(q_{k,0}(Z) - q_k(Z)\right) P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\left[\left(k(A; \tau, g) - k_0(A; \tau_0, g_0)\right)\right]\right]$$

$$= \left\langle (q_{k,0} - q_k), P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\left[k(\tau, g) - k_0(\tau_0, g_0)\right]\right\rangle$$

$$= \mathbb{E}\left[P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)}\left[\left(q_{k,0}(Z) - q_k(Z)\right)\right]\left(k(A; \tau, g) - k_0(A; \tau_0, g_0)\right)\right]$$

$$= \left\langle P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)}\left[q_{k,0} - q_k\right], \left(k(\tau, g) - k_0(\tau_0, g_0)\right)\right\rangle$$

$$\leq \min\left\{\left\|P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\left(k_0(\tau_0, g_0) - k(\tau, g)\right)\right\|_2 \left\|q_{k,0} - q_k\right\|_2, \ \left\|k_0(\tau_0, g_0) - k(\tau, g)\right\|_2 \left\|P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)}\left(q_{k,0} - q_k\right)\right\|_2\right\}$$

Term 2:

$$\mathbb{E}\left[\left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right)\left(\tau_0(Z; g_0) - \tau(Z; g)\right)\right]$$

$$= \mathbb{E}\left[\left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right) P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}\left[\tau_0(Z; g_0) - \tau(Z; g)\right]\right]$$

$$= \left\langle \left(q_{\tau,0}(q_k) - q_\tau(q_k)\right), P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}\left[\tau_0(g_0) - \tau(g)\right]\right\rangle$$

$$= \mathbb{E}\left[P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}\left[q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right]\left(\tau_0(Z; g_0) - \tau(Z; g)\right)\right]$$

$$= \left\langle P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}\left[q_{\tau,0}(q_k) - q_\tau(q_k)\right], \left(\tau_0(g_0) - \tau(g)\right)\right\rangle$$

$$\leq \min\left\{\left\|P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}\left(\tau_0(g_0) - \tau(g)\right)\right\|_2 \left\|q_{\tau,0}(q_k) - q_\tau(q_k)\right\|_2, \ \left\|\tau_0(g_0) - \tau(g)\right\|_2 \left\|P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}\left(q_{\tau,0}(q_k) - q_\tau(q_k)\right)\right\|_2\right\}$$

Term 3:

$$\mathbb{E}\left[\left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right)\left(g(Z) - g_0(Z)\right)\right]$$
$$= \left\langle \left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right), (g - g_0)\right\rangle$$
$$\leq \|(g - g_0)\|_2 \|\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\|_2$$

Joining these terms, we obtain

$$\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g)\right] - \theta_0$$
$$\leq \min\left\{\left\|P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\left(k_0(\tau_0, g_0) - k(\tau, g)\right)\right\|_2 \|q_{k,0} - q_k\|_2, \ \|k_0(\tau_0, g_0) - k(\tau, g)\|_2 \left\|P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)}\left(q_{k,0} - q_k\right)\right\|_2\right\}$$
$$+ \min\left\{\left\|P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}\left(\tau_0(g_0) - \tau(g)\right)\right\|_2 \|q_{\tau,0}(q_k) - q_\tau(q_k)\|_2, \ \|\tau_0(g_0) - \tau(g)\|_2 \left\|P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}\left(q_{\tau,0}(q_k) - q_\tau(q_k)\right)\right\|_2\right\}$$
$$+ \|(g - g_0)\|_2 \|\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\|_2.$$

$\square$

Proof of theorem 7. Let $\hat{\theta}_k = \frac{1}{|\mathcal{I}_j|}\sum_{i \in \mathcal{I}_j} m_3\left(O; \hat{k}^{(j)}, \hat{\tau}^{(j)}, \hat{g}^{(j)}, \hat{q}_k^{(j)}, \hat{q}_\tau^{(j)}, \hat{\alpha}_g^{(j)}\right)$. Also, let $H = (k, \tau, g, q_k, q_\tau, \alpha_g)$, and $H_0 = (k_0, \tau_0, g_0, q_{k,0}, q_{\tau,0}, \alpha_{g,0})$ as well as $\hat{H}^{(j)} = (\hat{k}^{(j)}, \hat{\tau}^{(j)}, \hat{g}^{(j)}, \hat{q}_k^{(j)}, \hat{q}_\tau^{(j)}, \hat{\alpha}_g^{(j)})$, so $\hat{\theta}_k = \mathbb{E}_{n,k}\left[m_3(O; \hat{H}^{(j)})\right]$. Expand the above to get

$$\hat{\theta}_k - \theta_0 = \mathbb{E}_{n,k}\left[m_3(O; \hat{H}^{(j)}) - \theta_0\right]$$
$$= \mathbb{E}_{n,k}\left[m_3(O; H_0) - \theta_0\right] + \mathbb{E}_{n,k}\left[m_3(O; \hat{H}^{(j)}) - m_3(O; H_0)\right]$$
$$= \mathbb{E}_{n,k}\left[m_3(O; H_0) - \theta_0\right] + \left(\mathbb{E}_{n,k} - \mathbb{E}\right)\left[m_3(O; \hat{H}^{(j)}) - m_3(O; H_0)\right] + \mathbb{E}\left[m_3(O; \hat{H}^{(j)}) - m_3(O; H_0)\right].$$

By assumption 6.2, $\|\hat{\eta}^{(j)} - \eta_0\|_2^2 = o_p(1)$ for any $\eta \in H$. It follows from simple algebra that there exists a universal constant such that

$$\mathbb{E}\left[\left(m_3(O; \hat{H}^{(j)}) - m_3(O; H_0)\right)^2\right] \leq c \sum_{\eta \in H} \|\hat{\eta}^{(j)} - \eta_0\|_2^2 = o_p(1).$$

Then, using the Markov inequality,

$$\left|\left(\mathbb{E}_{n,k} - \mathbb{E}\right)\left[m_3(O; \hat{H}^{(j)}) - m_3(O; H_0)\right]\right| = o_p\left(|\mathcal{I}_j|^{-1/2}\right) = o_p\left(n^{-1/2}\right).$$

Dealing with $\mathbb{E}\left[m_3(O; \hat{H}^{(j)}) - m_3(O; H_0)\right]$ requires assumptions on the rates on the compo-

nents of its error decomposition. By corollary 3,

$$\left| \mathbb{E}\left[m_3(O; \hat{H}^{(j)}) - m_3(O; H_0) \right] \right|$$

$$\leq \min\left\{ \left\| P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\left(k_0(\tau_0, g_0) - \hat{k}^{(j)}(\hat{\tau}^{(j)}, \hat{g}^{(j)}) \right) \right\|_2 \left\| q_{k,0} - \hat{q}_k^{(j)} \right\|_2 , \; \left\| k_0(\tau_0, g_0) - \hat{k}^{(j)}(\hat{\tau}^{(j)}, \hat{g}^{(j)}) \right\|_2 \left\| P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)}\left(q_{k,0} - \hat{q}_k^{(j)} \right) \right\|_2 \right\}$$

$$+ \min\left\{ \left\| P_{\mathcal{L}_2(W)}^{Z,\mathcal{T}}\left(\tau_0(g_0) - \hat{\tau}^{(j)}(\hat{g}^{(j)}) \right) \right\|_2 \left\| q_{\tau,0}(\hat{q}_k^{(j)}) - \hat{q}_\tau^{(j)}(\hat{q}_k^{(j)}) \right\|_2 , \; \left\| \tau_0(g_0) - \hat{\tau}^{(j)}(\hat{g}^{(j)}) \right\|_2 \left\| P_{Z,\mathcal{T}}^{\mathcal{L}_2(W)}\left(q_{\tau,0}(\hat{q}_k^{(j)}) - \hat{q}_\tau^{(j)}(\hat{q}_k^{(j)}) \right) \right\|_2 \right\}$$

$$+ \left\| \left(\hat{g}^{(j)} - g_0 \right) \right\|_2 \left\| \alpha_{g,0}(Z; \hat{q}_k^{(j)}, \hat{q}_\tau^{(j)}) - \hat{\alpha}_g^{(j)}(Z; \hat{q}_k^{(j)}, \hat{q}_\tau^{(j)}) \right\|_2 .$$

Using definition 6.4, the above can be written as

$$\left| \mathbb{E}\left[m_3(O; \hat{H}^{(j)}) - m_3(O; H_0) \right] \right|$$

$$\leq \min\left\{ \epsilon_{n,k,\mathcal{L}_2(Z)}\epsilon_{n,q_k}, \epsilon_{n,k}\epsilon_{n,q_k,\mathcal{K}} \right\} + \min\left\{ \epsilon_{n,\tau,\mathcal{L}_2(W)}\epsilon_{n,q_\tau}, \epsilon_{n,\tau}\epsilon_{n,q_\tau,\mathcal{T}} \right\} + \epsilon_{n,g}\epsilon_{n,\alpha_g} = o_p\left(n^{-1/2} \right),$$

where the final equality holds by the assumption on L2 rate conditions, assumption 6.2.2.

This simplifies the problem to

$$\hat{\theta}_k - \theta_0 = \mathbb{E}_{n,k}\left[m_3(O; H_0) - \theta_0 \right] + o_p\left(n^{-1/2} \right)$$

$$\sqrt{n}\left(\hat{\theta}_k - \theta_0 \right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(m_3(O_i; H_0) - \theta_0 \right) + o_p(1).$$

By the Central Limit Theorem, as $n \to \infty$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0 \right) \to N(0, \sigma_0^2), \qquad \qquad \sigma_0^2 = \mathbb{E}\left[\left(m_3(O; H_0) - \theta_0 \right)^2 \right].$$

By Slutsky's theorem, as long as $\hat{\sigma}_n^2 \to \sigma_0^2$ in probability for an estimator $\hat{\sigma}_n^2$ of σ_0^2, it holds that as $n \to \infty$,

$$\frac{\sqrt{n}}{\hat{\sigma}_n^2}\left(\hat{\theta}_n - \theta_0 \right) \to N(0, 1).$$

$\square$

11 Data Description

The sample consists of 1,983 individuals.

*Y***: Household net worth at 35 (Z9141400)**

Household net worth was top-coded at 600,000$ and bottom-coded at -300,000$. 7.0% of individuals were top-coded, 0.3% bottom-coded.

Figure 4: Histogram for *Y*
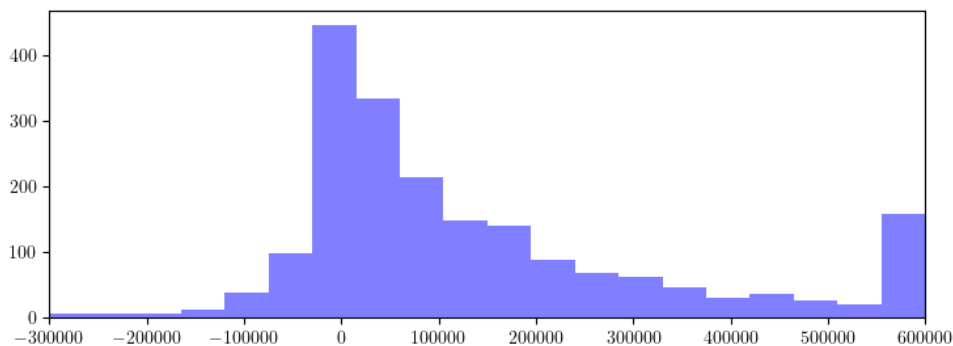
Notes: Distribution of household net worth.

*A***: Bachelor's degree obtained (Z9084400)**

If there is a date of obtaining a bachelor's degree (Z9084400 ≥ 0 or invalid skip -3), $A = 1$. 50.0% of individuals in the sample have obtained a BA degree.

*Z***: Pre-college ability measures**

Instruments are credit-weighted high-school GPAs in English (R9872000), Math (R9872200), Social Sciences (R9872300) and Life Sciences (R9872400), as well as the ASVAB percentile in each individual's respective age group.

*W***: Pre-college risky behaviour**

The proxies consist of risky behaviour dummies. They equal one when an individual has engaged in behaviour considered "risky" by age 17 or earlier if missing.

Table 4: Probability of engaging in risky behaviour *W* by 17 (in %)

	drink	smoke	mari-juana	run away	attack someone	sell drugs	destroy property	steal < 50$	steal > 50$
Pr	65.5	47.1	29.8	10.9	19.0	9.1	32.0	38.1	8.0

Figure 5: Histogram for high-school GPA
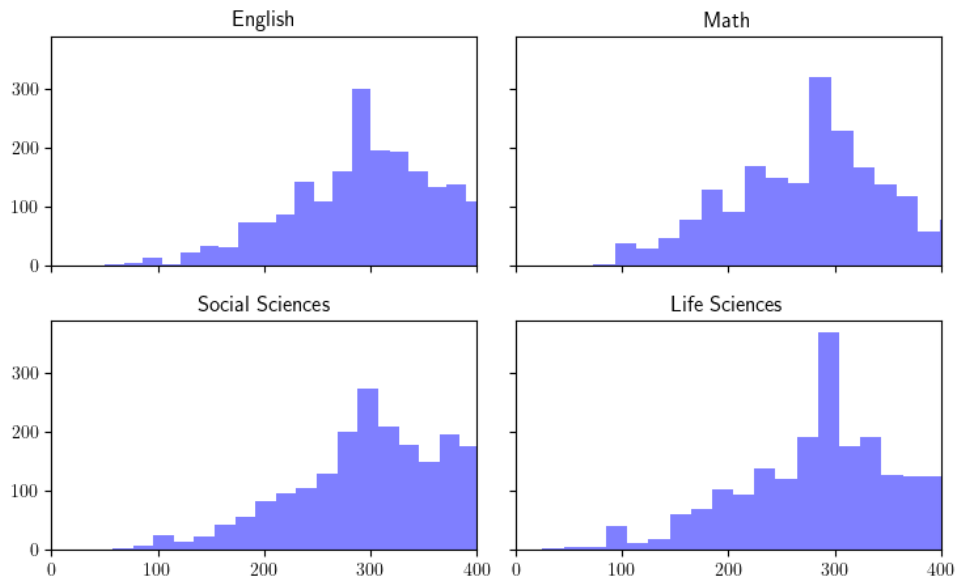
Notes: Distribution of household high-school GPAs.

X: **Individual, family, and regional covariates**

The proxies consist of risky behaviour dummies. They equal one when an individual has engaged in behaviour considered "risky" by age 17 or earlier if missing.

Figure 6: Histogram for ASVAB percentile in 3-month age cohort
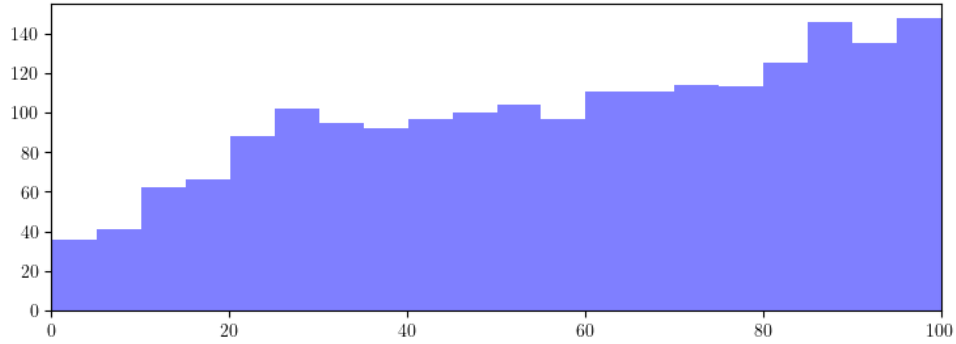
Notes: Distribution of ASVAB percentiles.

Table 5: Correlation matrix of pre-college risky behaviour W (in %)

	drink	smoke	mari-juana	run away	attack someone	sell drugs	destroy property	steal < 50$	steal > 50$
drink	100	49	41	18	15	21	24	28	14
smoke	49	100	49	21	20	28	27	33	17
marijuana	41	49	100	26	24	43	29	35	21
run away	18	21	26	100	27	22	20	17	23
attack	15	20	24	27	100	27	28	24	18
sell drugs	21	28	43	22	27	100	30	27	27
destroy property	24	27	29	20	28	30	100	39	21
steal < 50$	28	33	35	17	24	27	39	100	29
steal > 50$	14	17	21	23	18	27	21	29	100

Table 6: Covariates in first stage

variable	NLS97 basis	info	replace invalid by
hh net worth 1997	R1204700		35% quantile (8834.35)
male	R0536300	46.6%	
citizen birth non us	R1201300	2.3% (not US-born)	
citizen birth other	R1201300	13.2% (not det.)	
bio mom age first birth	R1200100		median (23)
bio mom age subject birth	R1200100		median (26)
bio mom educ	R1302500		median (13)
bio dad educ	R1302400		median (12)
relation parent figure	R1205300		both biological parents (1)
parent religious	R1486900		median (3)
n siblings	T6745900		median (2)
region north central at 17	R1200300 (i.a.)	27.8%	
region south at 17	R1200300 (i.a.)	36.3%	
region west at 17	R1200300 (i.a.)	21.6%	
urban at 17	R1217500 (i.a.)	69.8%	
other non rural at 17	R1217500 (i.a.)	0.3%	
non english home	R0551900	14.1%	0
poor english interview	R2394903	0.3%	0

Table 7: Additional covariates in outcome model

variable	NLS97 basis	info	replace invalid by
net worth age 20	Z9048900		35% quantile (2737.7)
gpa college	B0004600		400+ as 400
region north central at 35	U0001900 (i.a.)	24.5%	
region south at 35	U0001900 (i.a.)	38.7%	
region west at 35	U0001900 (i.a.)	23.2%	
urban at 35	U0015000 (i.a.)	81.9%	
other non rural at 35	U0015000 (i.a.)	0.4%	

Table 8: OLS coefficients key variables

	OLS	NC	IV	ICC
const	-32.36	-30.21	46.85	13.52
	(37.41)	(36.21)	(43.38)	(45.10)
T		27.76***		16.05**
		(4.82)		(7.37)
A	59.18***	30.90***	222.97***	125.15**
	(9.12)	(10.40)	(34.74)	(52.93)

Notes: The table contains estimates and their standard errors (in paran-theses) for β in the A row, and the linear parameter on T if used in the method from four estimators: Ordinary Least Squares (OLS), Proximal Learning (NC), Instrumental Variables (IV), and Instrumented Common Confounding (ICC). Asterisks indicate significance at the 1% (***), 5% (**) and 10% (*) level.

Table 9: OLS coefficients individual- and family-covariates

	OLS	NC	IV	ICC
hh net worth 1997	0.15***	0.16***	0.12***	0.14***
	(0.03)	(0.03)	(0.03)	(0.03)
net worth age 20	0.35***	0.36***	0.35***	0.36***
	(0.09)	(0.08)	(0.09)	(0.08)
gpa college	16.00***	11.58***	-6.01	2.86
	(3.44)	(4.00)	(6.19)	(6.94)
male	30.19***	24.33***	40.32***	32.06***
	(8.39)	(7.94)	(9.14)	(9.05)
n siblings	-6.29***	-6.78***	-6.08**	-6.47***
	(1.73)	(2.16)	(2.37)	(2.21)
citizen birth non us	-5.50	1.61	-16.48	-7.42
	(27.65)	(28.22)	(30.87)	(29.11)
citizen birth other	34.53***	37.24***	29.66**	33.18**
	(13.24)	(12.88)	(14.09)	(13.27)
bio mom age first birth	2.28**	2.60**	1.11	1.80
	(1.00)	(1.03)	(1.15)	(1.13)
bio mom age subject birth	-0.08	-0.06	-0.01	0.00
	(0.55)	(0.57)	(0.62)	(0.58)
bio mom educ	2.91	3.21*	-0.16	1.53
	(1.95)	(1.77)	(2.04)	(2.03)
bio dad educ	1.84	3.22*	-2.52	0.38
	(1.84)	(1.71)	(2.06)	(2.32)
relation parent figure	-8.81***	-10.29***	-5.41**	-7.78***
	(2.22)	(2.44)	(2.75)	(2.82)
parent religious	-0.03	-0.03	-0.04	-0.03
	(0.03)	(0.03)	(0.03)	(0.03)
non english home	-10.64	-12.37	-9.47	-11.23
	(13.01)	(13.20)	(14.40)	(13.43)
poor english interview	75.41	56.96	78.91	65.72
	(112.29)	(78.33)	(85.58)	(79.79)

Notes: The table contains estimates and their standard errors (in parantheses) for slope coefficients on individual and family covariates using four estimators: Ordinary Least Squares (OLS), Proximal Learning (NC), Instrumental Variables (IV), and Instrumented Common Confounding (ICC). Asterisks indicate significance at the 1% (***), 5% (**) and 10% (*) level.

Table 10: Slope coefficients regional covariates

	OLS	NC	IV	ICC
urban 17	1.08	-1.48	5.09	1.59
	(10.04)	(9.68)	(10.64)	(10.01)
urban 35	-30.42***	-28.21***	-38.53***	-32.79***
	(11.48)	(10.89)	(12.00)	(11.47)
other non rural 17	24.13	19.02	42.62	30.34
	(23.31)	(24.07)	(26.55)	(25.25)
other non rural 35	-92.49	-82.51	-48.20	-66.35
	(72.08)	(66.52)	(73.14)	(68.57)
region north central 17	43.59*	42.51**	38.27*	40.92**
	(22.91)	(20.27)	(22.13)	(20.68)
region north central 35	-28.57	-27.57	-11.03	-19.45
	(24.64)	(21.07)	(23.28)	(22.11)
region south 17	16.45	14.74	8.67	11.96
	(18.82)	(18.22)	(19.92)	(18.66)
region south 35	-22.18	-20.87	-1.67	-11.54
	(19.98)	(18.40)	(20.49)	(19.65)
region west 17	27.39	25.91	24.62	25.23
	(22.03)	(19.97)	(21.79)	(20.34)
region west 35	16.58	15.37	34.70	24.20
	(22.98)	(20.06)	(22.19)	(21.22)

Notes: The table contains estimates and their standard errors (in parantheses) for slope coefficients on regional covariates using four estimators: Ordinary Least Squares (OLS), Proximal Learning (NC), Instrumental Variables (IV), and Instrumented Common Confounding (ICC). Asterisks indicate significance at the 1% (***), 5% (**) and 10% (*) level.

Table 11: Slope coefficients risky behaviour proxies

	OLS	IV
ever drank 17	14.44	12.56
	(10.10)	(10.63)
ever smoked 17	-2.40	7.39
	(10.01)	(10.90)
ever marijuana 17	-9.03	-0.38
	(10.90)	(12.10)
ever ran away 17	-8.52	-12.39
	(12.91)	(14.74)
ever attack 17	-1.71	15.39
	(10.46)	(12.54)
ever sell drugs 17	-22.44	-19.73
	(15.87)	(17.27)
ever destroy property 17	3.86	2.57
	(9.89)	(10.59)
ever steal bit 17	-1.29	-4.15
	(9.50)	(10.30)
ever steal lot 17	-20.98	-11.98
	(15.45)	(17.12)

Notes: The table contains estimates and their standard errors (in parantheses) for slope coefficients on risky behaviour proxies using four estimators: Ordinary Least Squares (OLS), Proximal Learning (NC), Instrumental Variables (IV), and Instrumented Common Confounding (ICC). Asterisks indicate significance at the 1% (***), 5% (**) and 10% (*) level.