Auxiliary IV Estimation for Nonlinear Models*

Riccardo D'Adamo‡ Martin Weidner§ Frank Windmeijer¶

June 6, 2023

Abstract

This paper suggests a new instrumental variable (IV) estimator for non-linear models with endogenous covariates. We consider the estimation of the regression coefficients on the endogenous variables based on the following criterion: If the IVs are added as auxiliary regressors to the model, then we want their estimated coefficients to be equal or close to zero. This method is quite intuitive as it formalizes the idea that the IVs should be excluded variables that do not have any direct explanatory power for the outcome. Whilst this estimator is not consistent in general, we derive that it is consistent when the coefficients of the endogenous variables are equal to zero, and that it is locally sign consistent. This is confirmed by some Monte Carlo simulations. The usefulness of the estimator is further highlighted in two empirical examples. The focus of the paper is mainly on the binary choice model, but the results extend to other non-linear models.

1 Introduction

Instrumental variables (IVs) are an essential tool to estimate causal relationships from observational data. The underlying idea has been around for nearly a century (going back to the appendix in Wright 1928), and the "credibility revolution" in empirical economics has raised attention to IV methods even further in the past few decades (see e.g. Angrist and Pischke 2010). Accordingly, there is a very large literature on the subject, but the

*This research was supported by the European Research Council grant ERC-2018-CoG-819086-PANEDA.

‡Aarhus University, r.dadamo@econ.au.dk

§University of Oxford, martin.weidner@economics.ox.ac.uk

¶University of Oxford, frank.windmeijer@stats.ox.ac.uk

majority of both applied and theoretical work focuses on estimating linear regression models. IV estimation of non-linear models is a challenging problem, and, despite a lot of work on the subject (see references below), there is still room for new methods and ideas.

The aim of this paper is to estimate non-linear models with endogenous covariates when appropriate IVs are available. However, to explain our estimation approach in the simplest possible setting, consider a linear regression model for a scalar outcomes Y_i, with a vector of potentially endogenous covariates X_i, and a vector of instruments Z_i, observed for units $i = 1, \dots, n$. We are interested in the effect of X_i on Y_i, parameterized by the vector β. There are many ways to construct an IV estimator in a linear model, and, at least for the case of a limited number of strong instruments, they are all essentially equivalent (up to some choice of appropriate weight matrix). One of those ways is as follows: Let $\widehat{\gamma}(\beta)$ be the ordinary least squares (OLS) estimator obtained by regressing Z_i on the residuals $Y_i - X_i' \beta$, and let $\widehat{\beta}$ be obtained by minimizing the objective function $\widehat{\gamma}'(\beta)\, \Omega\, \widehat{\gamma}(\beta)$. Here, Ω is a symmetric positive definite weight matrix. For example, if we set $\Omega = \sum_{i=1}^N Z_i\, Z_i'$, then, under standard regularity conditions, it is easy to verify that $\widehat{\beta}$ is equal to the two-stage least squares (2SLS) estimator.[1]

This procedure of obtaining the 2SLS estimator is quite intuitive: We choose β such that Z_i has no explanatory power for the residuals $Y_i - X_i' \beta$, or equivalently, such that the regression coefficient of Z_i on $Y_i - X_i' \beta$ is (close to) zero. This is one way of formalizing what is meant by the instrument being an excluded variable.

Depending on the underlying model specification, we might not want to obtain $\widehat{\gamma}(\beta)$ by OLS. For example, Chernozhukov and Hansen (2006) apply this estimation approach for quantile regressions with endogeneity, that is, $\widehat{\gamma}(\beta)$ is obtained by a quantile regression (Koenker and Bassett 1978) of Z_i on $Y_i - X_i' \beta$. Similarly, Lee, Moon and Weidner (2012) estimate panel regression models with endogeneity and unobserved factors, and therefore obtain $\widehat{\gamma}(\beta)$ by a panel regression with unobserved factors (Pesaran 2006; Bai 2009). Those ideas are combined by Harding and Lamarche (2014) who obtain $\widehat{\gamma}(\beta)$ by a quantile regression that also controls for unobserved factors.

In all those papers, the relation between Y_i and X_i is still linear. In the current paper, we generalize this estimation approach to models where the relation between Y_i and X_i is non-linear. Our leading example is the binary choice model $Y_i = \mathbb{1}\{X_i' \beta + U_i \geq 0\}$, where the distribution of the unobserved error U_i is assumed to be known (e.g. a logit or

[1] This is a representation of 2SLS as a minimum-distance estimator. Windmeijer (2019) shows that 2SLS can be expressed in a different way as a minimum-distance estimator as well.

probit model), and U_i is independent of Z_i, but may be correlated with X_i.

In this model, if X_i were exogenous, then we would simply use the maximum likelihood estimator (MLE) to estimate β. For the case of endogenous covariates, it therefore seems natural to obtain $\widehat{\gamma}(\beta)$ as the MLE of the model $Y_i = \mathbb{1}\{X_i'\beta + Z_i'\gamma + U_i \geq 0\}$, where β is fixed, and the likelihood function is only maximized over γ. The estimator for β is then obtained by minimizing $\widehat{\gamma}'(\beta)\,\Omega\,\widehat{\gamma}(\beta)$, as before. We denote the resulting estimator for β the Auxiliary IV (AIV) estimator, because the instrument Z_i is included as an auxiliary regressor in the maximum likelihood estimation.

We find this AIV estimator very natural and intuitive, and the goal of this paper is to show that it has interesting theoretical properties and is useful in practice. However, the problem of IV estimation of non-linear models is too complicated to expect that the AIV estimator is a miracle solution that always works well. In particular, under the model assumptions imposed so far, the AIV will generally not be consistent for the true parameter value for β (as $n \to \infty$). This is because the estimator $\widehat{\gamma}(\beta)$ is obtained by maximizing a misspecified likelihood function: When we write down the likelihood for the model $Y_i = \mathbb{1}\{X_i'\beta + Z_i'\gamma + U_i \geq 0\}$, we use the distribution of U_i conditional on Z_i, which is assumed to be known by the model assumptions, but one should really use the distribution of U_i conditional X_i and Z_i. The latter is, however unknown without further assumptions on the data generating process for the endogenous X_i.

The main reason why we think that the AIV estimator is useful despite being inconsistent in general is the following: If $\beta = 0$ (or more precisely, if the coefficients on the endogenous components of X_i are zero), then the AIV estimator *is* consistent as $n \to \infty$, and it also typically estimates the sign of β correctly within a neighborhood of β. This local sign consistency is a very useful property in empirical applications, where it is often a primary concern whether a coefficient is different from zero, and what the sign of a coefficient is. Further, the AIV estimator is a plausible estimator for β that can be constructed without making any assumptions on the data generating process for X_i. The endogenous regressors can be discrete or continuous, and apart from regularity conditions, can be arbitrarily distributed and arbitrarily correlated with U_i. This should be contrasted with other simple IV estimators for non-linear models like the control function estimator (Rivers and Vuong 1988) or the joint MLE that also fully parameterizes the distribution of X_i. Such distributional assumptions are seldom justified by economic theory, and it is well known that maximum likelihood estimators of bivariate models can be very sensitive to misspecification of the error distribution (Little, 1985; Monfardini and Radice, 2008)

We are therefore confident that the AIV estimator is a useful addition to the toolbox of applied researchers, which should be reported alongside other estimation approaches that have complementary properties, as illustrated by the empirical applications in this paper.

As already mentioned above, there is a large existing literature on IV estimation in both linear and non-linear models. General non-parametric identification results are discussed, for example, in Imbens and Newey (2009), Chesher (2010), and Chesher and Rosen (2017).

Newey (1986) presents a weighted IV estimator for continuous endogenous regressors that requires estimation of the density of the exogenous regressors and instruments, and assumes linearity of the first-stage equation. Yildiz (2013) proposes a matching estimator that is $\sqrt{n}$-consistent for the coefficient of the single binary endogenous variable under non-parametric restrictions on the distribution of the unobservables, but relies on parametric specification of the functional form for the first-stage equation (e.g. a linear index specification). Han and Lee (2019) consider estimation of generalized bivariate probit models under a parametric copula assumption for the errors. The validity of their proposed procedure does not rely on knowledge of the marginal distribution of the errors in the structural and first-stage equations, but requires a parametric specification of the functional form of the first-stage equation.[2] Our proposed estimator assumes knowledge of the distribution of the error in the structural equation but does not impose any functional form or distributional assumption on the first-stage equation. As a result, our proposed estimator has complementary properties to those mentioned above.

Abrevaya, Hausman and Khan (2010) provide a consistent test for the relevance and sign of the endogenous regressor under no parametric assumptions on the distribution of the errors. Their test is based on a version of Kendall's τ-statistic that uses fitted values from the first-stage equation. Unlike Abrevaya, Hausman and Khan (2010), the validity of the test of regressor relevance based on the AIV estimator does not rely on parametric assumptions on the functional form of the first-stage.

Mu and Zhang (2018) propose an estimator for triangular binary choice models with a binary endogenous regressor based on maximum score Manski (1985). Their proposal relies on the existence of continuous exogenous regressors with large support, in the spirit of Lewbel (2000). Their procedure does not require parametric specification of the dis-

[2] Han and Lee (2019) also discuss identification in bivariate probit models in the absence of excluded instruments. See also Mourifié and Méango (2014) and Han and Vytlacil (2017).

tribution of unobservables or the endogenous regressor, but leads to rates of convergence that can be considerably slower than $\sqrt{n}$.

Bhattacharya, Shaikh and Vytlacil (2012) show 2SLS with a binary outcome and binary endogenous regressor correctly estimates the sign of the average treatment effect. This property of 2SLS however is not guaranteed in the presence of additional exogenous covariates. Our simulations suggest that, unlike 2SLS, the AIV estimator's sign-consistency property is robust to the inclusion of additional regressors.

Our results for the AIV estimator of consistency at $\beta = 0$ and the local sign consistency generalise the results in the epidemiology literature of Dai and Zhang (2015), who show this result for the logit model with a continuous endogenous regressor when it is replaced by its first-stage linear IV prediction.

In the following, we first introduce the model assumptions and AIV estimator in Section 2, for the case where the only unknown parameters are the slope coefficients in a single index. The large sample properties of the estimator are then studied in Section 3. Generalizations to models with additional parameters are discussed in Section 4. Monte Carlo results and empirical applications are presented in Section 5 and 6 respectively. Finally, Section 7 concludes.

2 Model and Auxiliary IV estimator

2.1 Model

For each unit $i = 1, \ldots, n$ we observe a scalar outcome $Y_i \in \mathcal{Y}$, a vector of covariates X_i, and a vector of instrumental variables Z_i. In practice, often only a subset of the covariates are considered to be endogenous, in which case the exogenous covariates are included in Z_i. We denote the dimension of X_i and Z_i by $k_x \in \{1, 2, \ldots\}$ and $k_z \in \{1, 2, \ldots\}$, respectively.

Assumption 1 (Model).

(i) The outcomes Y_i are generated from the latent variable model

$$Y_i = g(\omega_{0,i}, U_i), \qquad \omega_{0,i} := X_i' \beta_0,$$

where $U_i \in \mathbb{R}$ are unobserved random variables, the function $g(\cdot, \cdot)$ is known, and β_0 are vectors of unknown parameters.

(ii) *The distribution of U_i is independent of Z_i, and U_i has known cumulative distribution function $F_U(\cdot)$.*

(iii) *(X_i, Z_i, U_i) are independent and identically distributed across $i = 1, \ldots, n$.*

For example, for a binary choice model the function $g(\cdot, \cdot)$ in Assumption 1(i) is given by $g(\omega, u) = \mathbb{1}\{\omega + u \geq 0\}$ and we have

$$Y_i = \mathbb{1}\{X_i' \beta_0 + U_i \geq 0\}.$$

In particular, for a binary choice probit model we choose the distribution of U_i to be standard normal, or $F_U(\cdot) = \Phi(\cdot)$ in Assumption 1(ii), with $\Phi(\cdot)$ denoting the standard normal cdf. Notice that Assumption 1(ii) imposes independence between the unobserved error U_i and the instrument Z_i, but the covariate X_i may be correlated with U_i. Finally, Assumption 1(iii) imposes cross-sectional sampling.

The binary choice probit model is our leading example that will be used throughout most of the paper. However, Assumption 1 also covers, amongst others, a Poisson model. Section 4 discusses more general models where $Y_i = g(\omega_{0,i}, U_i)$ is replaced by $Y_i = g(\omega_{0,i}, W_i, U_i, \alpha_0)$, with additional unknown parameters α_0 and additional exogenous covariates W_i. That extension is important to cover models that feature additional unknown parameters beyond the regression coefficients β_0, for example, Tobit models, ordered choice models, or multinomial choice models. However, to present our main idea and results as clearly as possible we find it convenient to focus on the simpler model structure in Assumption 1 first, which covers the binary choice model as our leading example.

For the model described by Assumption 1, let $\ell\left(y \,\middle|\, \omega\right)$ denote the log-likelihood of observing $Y_i = y$ conditional on $\omega_{0,i} = \omega \in \mathbb{R}$, treating U_i and $\omega_{0,i}$ as independent. For discrete Y_i we have

$$\ell\left(y \,\middle|\, \omega\right) = \log \Pr\{y = g(\omega, U_i)\},$$

where the probability is evaluated according to the cdf $F_U(\cdot)$. For all our theoretical results below we will assume that the log-likelihood is strictly concave and continuously differentiable in ω. This is, of course, satisifed for the binary choice probit model where $\ell\left(y \,\middle|\, \omega\right) = y \log \Phi(\omega) + (1 - y) \log[1 - \Phi(\omega)]$.

2.2 AIV estimator

If Assumption 1 holds with $Z_i = X_i$, then X_i is exogenous and the most natural estimator for β in the model described above is given by the maximum likelihood estimator (MLE)

$$\widehat{\beta}_{\text{MLE}} = \underset{\beta}{\text{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, X_i'\beta\right).$$

However, if $Z_i \neq X_i$ and (some of) the covariates X_i are endogenous, then $\widehat{\beta}_{\text{MLE}}$ is generally not a good estimator anymore. Some estimation strategy that makes use of the instrumental variables Z_i is required in that case. The auxiliary IV estimator $\widehat{\beta}_{\text{AIV}}$ that we propose is defined by

$$\begin{aligned} \widehat{\gamma}(\beta) &= \underset{\gamma \in \mathcal{C}}{\text{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right), \\ \widehat{\beta}_{\text{AIV}} &\in \underset{\beta \in \mathcal{B}}{\text{argmin}} \, \|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}, \end{aligned} \tag{1}$$

where $\mathcal{C} \subset \mathbb{R}^{k_z}$ and $\mathcal{B} \subset \mathbb{R}^{k_x}$ are compact sets, and $\|\gamma\|_\Omega^2 = \gamma' \Omega \gamma$ is a quadratic distance measure for vectors $\gamma \in \mathbb{R}^{k_z}$, parameterized by a positive definite $k_z \times k_z$ weight matrix $\Omega = \Omega_{n,\beta}$, which might be stochastic and might depend on β. If we choose Ω equal to the identity matrix, then $\|\cdot\|_\Omega$ is simply the Euclidean norm. But having the flexibility to choose more general Ω is useful, for example, by choosing $\Omega = \frac{1}{n}\sum_i Z_i Z_i'$ the estimator $\widehat{\beta}_{\text{AIV}}$ remains unchanged under the transformation $Z_i \mapsto Z_i A$, for any invertible $k_z \times k_z$ matrix A.

We introduce the compact sets $\mathcal{C}$ and $\mathcal{B}$ for technical reasons. In our practical implementation we assume that the boundedness conditions imposed through $\mathcal{C}$ and $\mathcal{B}$ are non-binding, that is, in practice we implement $\widehat{\beta}_{\text{AIV}}$ with $\mathcal{C} = \mathbb{R}^{k_z}$ and and $\mathcal{B} = \mathbb{R}^{k_x}$.

For the special case of all regressors known to be exogenous, $Z_i = X_i$, we have $\widehat{\gamma}(\beta) = \widehat{\beta}_{\text{MLE}} - \beta$, and therefore $\widehat{\beta}_{\text{AIV}} = \widehat{\beta}_{\text{MLE}}$. Also, for the linear regression model, $Y_i = X_i'\beta_0 + U_i$, with normal errors $U_i \sim \mathcal{N}(0, \sigma_0^2)$ and $\Omega = \frac{1}{n}\sum_i Z_i Z_i'$ one can easily show that $\widehat{\beta}_{\text{AIV}} = \widehat{\beta}_{\text{2SLS}}$, as long as the boundedness conditions imposed through $\mathcal{C}$ and $\mathcal{B}$ are non-binding.

The idea underlying the IV estimator $\widehat{\beta}_{\text{AIV}}$ is as follows. We include the instruments Z_i as auxiliary regressors in the model and for fixed β we maximize the corresponding log-likelihood $\ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)$ only over the parameters γ that correspond to the exogenous variables Z_i. Intuitively, the instruments Z_i should be excluded variables and their coefficient estimates $\widehat{\gamma}(\beta)$ are therefore expected to be close to zero whenever β is close

to the true value β_0. Following that intuition we therefore obtain $\widehat{\beta}_{\text{AIV}}$ by minimizing the distance between $\widehat{\gamma}(\beta)$ and zero.

The idea of using instrumental variables as auxiliary regressors and then minimizing their coefficients to find the parameters of interest has previously been used in other contexts. In a quantile regression setting, this method was proposed by Chernozhukov and Hansen (2006). To deal with endogeneity in panel regressions with interactive fixed effects and for the purpose of demand estimation the method was used in Lee, Moon and Weidner (2012) and Moon, Shum and Weidner (2018). However, none of those existing papers consider the type of non-linear models with endogeneity that are the focus here. The IV estimator in (1) and our theoretical results below are novel in that context.

An interesting alternative characterization of the objective function $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}$ for $\widehat{\beta}_{\text{AIV}}$ is provided by the following lemma.

Lemma 1. *Let $\beta \in \mathbb{R}^{k_x}$. Let $W_{n,\beta} \in \mathbb{R}^{k_z \times k_z}$ be symmetric and positive definite. Assume that the log-likelihood $\ell\left(y \,\middle|\, \omega\right)$ is strictly concave and twice continuously differentiable in $\omega \in \mathbb{R}$, and that the maximizer $\widehat{\gamma}(\beta)$ in (1) is well-defined. Define the $k_z \times k_z$ matrix*[3]

$$H_n(\beta,\gamma) := \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2} Z_i\, Z_i'.$$

Then, there exists $\gamma_*(\beta) \in \mathbb{R}^{k_z}$ such that for

$$\Omega_{n,\beta} = H_n(\beta,\gamma_*(\beta))\; W_{n,\beta}\; H_n(\beta,\gamma_*(\beta)) \tag{2}$$

we have

$$\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}} = \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell\left(Y_i \,\middle|\, X_i'\beta\right)}{\partial \omega} Z_i \right\|_{W_{n,\beta}}.$$

The lemma provides an alternative characterization for the objective function $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}$ that is used to define our IV estimator $\widehat{\beta}_{\text{AIV}}$ in (1). For matrices $\Omega_{n,\beta}$ and $W_{n,\beta}$ satisfying the relation (2), we can use the lemma to express $\widehat{\beta}_{\text{AIV}}$ as

$$\widehat{\beta}_{\text{AIV}} \in \operatorname*{argmin}_{\beta \in \mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell\left(Y_i \,\middle|\, X_i'\beta\right)}{\partial \omega} Z_i \right\|_{W_{n,\beta}}. \tag{3}$$

[3] We use the following notation

$$\frac{\partial^q \ell\left(Y_i \,\middle|\, a_i\right)}{\partial \omega^q} := \left.\frac{\partial^q \ell\left(Y_i \,\middle|\, \omega\right)}{\partial \omega^q}\right|_{\omega = a_i}.$$

The researcher could choose the weight matrix $W_{n,\beta}$ (e.g. a fixed matrix independent of β) and use (3) to compute $\widehat{\beta}_{\text{AIV}}$. In that case, (1) provides an alternative characterization of the same $\widehat{\beta}_{\text{AIV}}$ as long as (2) holds. Or the researcher could choose the weight matrix Ω_β (e.g. a fixed matrix independent of β). Then, if $H_n(\beta, \gamma_*(\beta))$ is invertible, (3) provides an alternative characterization of the same $\widehat{\beta}_{\text{AIV}}$ as long as $W_{n,\beta} = [H_n(\beta, \gamma_*(\beta))]^{-1} \, \Omega_{n,\beta} \, [H_n(\beta, \gamma_*(\beta))]^{-1}$.

Furthermore, for the exactly identified case, $k_z = k_x$, if a solution $\widehat{\beta}_{\text{AIV}}$ of the method of moment equations

$$\sum_{i=1}^{n} \frac{\partial \ell \left(Y_i \,\middle|\, X_i' \, \widehat{\beta}_{\text{AIV}} \right)}{\partial \omega} Z_i = 0 \tag{4}$$

exists, then that solution also solves (3). Our assumptions in Section 3 guarantee existence of a solution to (4) for $k_z = k_x$ in large samples. Notice that (4) generalizes the first order condition of the MLE by replacing X_i with Z_i. While (4) is conveniently simple, we prefer the more general characterizations (1) and (3) of the estimator since they are applicable to the overidentified case, $k_z > k_x$, as well.

For both (1) and (3), the objective function for the minimization over β may not be convex. For computation we refer to Section 4. There we show that if only a single regressor is endogenous, then the "outer loop" optimization over β in (1) can be transformed into a one-dimensional problem (for which a grid search is computationally feasible), while the "inner loop" optimization over γ in (1) always remains a convex problem as long as the log-likelihood is concave.

3 Asymptotic results for the IV estimator

We have argued in the last section that the AIV estimator is a quite intuitive and plausible estimator to consider. However, IV estimation in non-linear models is a challenging problem and our relatively simple estimator $\widehat{\beta}_{\text{AIV}}$ does not miraculously fully solve this. Indeed, under the assumptions imposed so far, the IV estimator $\widehat{\beta}_{\text{AIV}}$ is *not* consistent for β_0 in general. Nevertheless, we believe that the estimator $\widehat{\beta}_{\text{AIV}}$ is a useful element in the toolbox of nonlinear IV estimation, and the purpose of the current section is to demonstrate this by deriving some asymptotic properties of $\widehat{\beta}_{\text{AIV}}$. To show consistency and asymptotic normality of $\widehat{\beta}_{\text{AIV}}$ we impose the following additional assumption.

Assumption 2 (Exogeneity of $X_i' \, \beta_0$). *U_i is independent of $(X_i' \beta_0, Z_i)$.*

Assumption 2 is satisfied if for every $k = 1, \ldots, k_x$ we *either* have $\beta_{0,k} = 0$ *or* $X_{i,k}$ is exogenous. Thus, endogenous regressors are allowed for here, as long as the corresponding coefficient is zero. Indeed, we are particularly interested in cases where some of the covariates $X_{i,k}$ are endogenous and the corresponding coefficients $\beta_{0,k}$ are close to zero, but the researcher may not know that the coefficients are close to zero. Those are the cases where the estimator $\widehat{\beta}_{\text{AIV}}$ will be most useful, either to formally test the null hypothesis $H_0 : \beta_{0,k} = 0$, or to simply report and interpret $\widehat{\beta}_{\text{AIV}}$ in a table with multiple other estimators that have complementary properties.

In subsections 3.1 we derive consistency and asymptotic normality of $\widehat{\beta}_{\text{AIV}}$ under Assumption 2. In subsection 3.2 we do not impose Assumption 2 strictly, but instead show that for endogenous $X_{i,k}$ we obtain local sign consistency for $\widehat{\beta}_{\text{AIV},k}$ in a neighborhood around $\beta_{0,k} = 0$.

3.1 Consistency and asymptotic normality

In addition to the Assumptions 1 and 2 imposed so far, we also require some more technical regularity conditions. For this purpose we introduce the matrices

$$G_n(\beta,\gamma) := \frac{1}{n}\sum_{i=1}^{n} Z_i X_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2}, \quad H_n(\beta,\gamma) := \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2},$$
$$G(\beta,\gamma) := \mathbb{E}\left[Z_i X_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2}\right], \quad H(\beta,\gamma) := \mathbb{E}\left[Z_i Z_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2}\right], \tag{5}$$

and the score function for γ:

$$S_n(\beta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} Z_i \frac{\partial \ell\left(Y_i \mid X_i'\beta + Z_i'\gamma\right)}{\partial \omega}. \tag{6}$$

Assumption 3 (Regularity conditions).

(i) The parameter sets $\mathcal{B}$ and $\mathcal{C}$ are compact. $\mathcal{B}$ contains β_0 as an interior point. $\mathcal{C}$ contains 0 as an interior point.

(ii) For all possible outcomes y, the log-likelihood function $\ell\left(y \,\middle|\, \omega\right)$ is strictly concave in $\omega \in \mathbb{R}$. Furthermore, $\ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)$ is three times continuously differentiable in (β,γ) with derivatives that in expectation are bounded for all $(\beta,\gamma) \in (\mathcal{B},\mathcal{C})$.

(iii) $\sup_{\beta\in\mathcal{B}}\sup_{\gamma\in\mathcal{C}}\|G_n(\beta,\gamma)-G(\beta,\gamma)\|=o_P(1)$, $\quad\sup_{\beta\in\mathcal{B}}\sup_{\gamma\in\mathcal{C}}\|H_n(\beta,\gamma)-H(\beta,\gamma)\|=o_P(1)$.

(iv) *For all* $(\beta,\gamma)\in(\mathcal{B},\mathcal{C})$, $H(\beta,\gamma)$ *has full rank* k_z *and* $G(\beta,0)$ *has full rank* k_x.

(v) *The symmetric matrix* $\Omega_{n,\beta}$ *is a twice continuously differentiable function in* β, *and there exists a constant* $c>0$ *such that with probability approaching one we have* $\Omega_{n,\beta}\geq c$ *for all* $\beta\in\mathcal{B}$. *Furthermore,* $\sup_{\beta\in\mathcal{B}}\|\Omega_{n,\beta}-\Omega_\beta\|=o_p(1)$ *for some non-random symmetric matrix* Ω_β *which is positive-definite for all* $\beta\in\mathcal{B}$.

Before we discuss these assumptions we first state our main consistency theorem.

Theorem 1. *Let Assumption 1, 2, 3 hold. Then we have* $\widehat{\beta}_{\text{AIV}}=\beta_0+o_P(1)$, *as* $n\to\infty$.

Assumptions 3(i) is a standard technical regularity condition that demands the parameter sets to be compact while also containing the true parameter values – notice that 0 is the "true value" for γ. Assumption 3(ii) demands the log-likelihood to be strictly concave and sufficiently smooth. Assumption 3(iii) is a uniform convergence requirement for the second derivatives of the sample likelihood function. Classic primitive conditions for uniform convergence through dominance conditions are satisfied under the smoothness assumptions in Assumption 3(ii) whenever $\mathbb{E}\left[\|Z_i'X_i\|\right]<\infty$ and $\mathbb{E}\left[\|Z_i'Z_i\|\right]<\infty$. Assumption 3(v) is a standard regularity condition on the weight matrix $\Omega_{n,\beta}$.

In Assumption 3(iv), the condition on $H(\beta,0)$ is a generalized non-collinearity condition on the instruments Z_i, while the condition on $G(\beta,0)$ is a generalized relevance condition on the instruments — if the definition of H and G in (5) would not contain $\partial^2\ell/\partial\omega^2$, then these would be the standard non-collinearity and relevance conditions. If one only wanted to show local consistency for $\mathcal{B}$ being a small neighborhood around β_0, then it would be sufficient to impose Assumption 3(iv) at β_0 only.

Theorem 2. *Suppose that Assumptions 1, 2, and 3 hold. Furthermore, assume that* $\sqrt{n}\,S_n(\beta_0,0)\xrightarrow{d}\mathcal{N}(0,\Sigma)$ *with* $\Sigma:=\operatorname{Var}\left[Z_i\frac{\partial\ell\left(Y_i\,|\,X_i'\beta\right)}{\partial\omega}\right]$. *Then,*

$$\sqrt{n}\left(\widehat{\beta}_{\text{AIV}}-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\,(G'\,W\,G)^{-1}\,G'\,W\,\Sigma\,W\,G\,(G'\,W\,G)^{-1}\right),$$

where $G:=G(\beta_0,0)$ *and* $W:=H^{-1}\,\Omega\,H^{-1}$, *with* $\Omega=\Omega_{\beta_0}$ *and* $H:=H(\beta_0,0)$.

Asymptotic normality of the score $S_n(\beta_0,0)$ can be shown using the Lindeberg-Lévy central limit theorem under the moment bound $\mathbb{E}\left[\|Z_i'Z_i\|\right]<\infty$. Apart from that, the assumptions of Theorem 2 are identical to those of Theorem 1. From the asymptotic

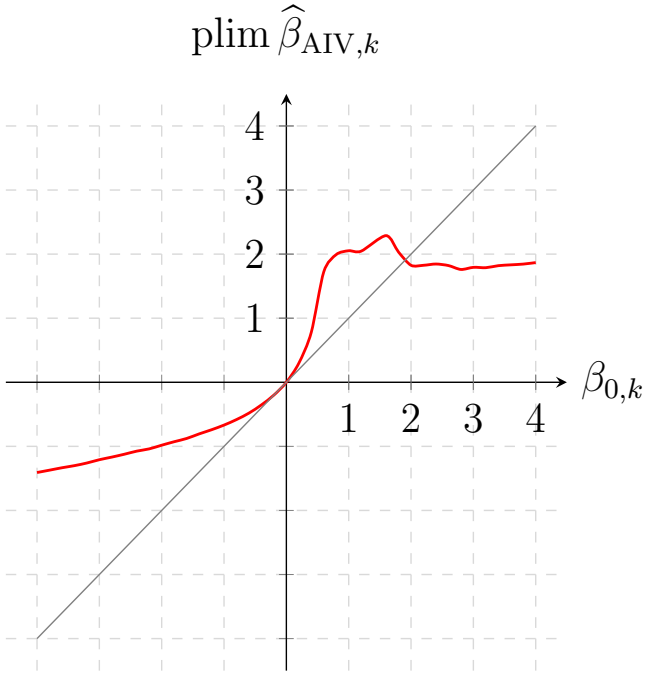

Figure 1: Probability limit of $\widehat{\beta}_{\text{AIV}}$ as a function of β_0.

variance formula of the AIV estimator one can deduce the optimal weighting matrix $\Omega^* = H\Sigma^{-1}H$ under which $\text{AVar}(\sqrt{\text{n}}\,\widehat{\beta}_{\text{AIV}}) = (G'\Sigma^{-1}G)^{-1}$. While continuously-updating or feasible two-step procedures would be asymptotically efficient, we find that they bring negligible gains in our simulations compared to the simple choice $\Omega_{n,\beta} = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'$, which is the one we recommend.

3.2 Local sign consistency

In this section we consider the case where all regressors are exogenous, except for a single endogenous regressor $X_{i,k}$. We are interested in how the probability limit of the corresponding component $\widehat{\beta}_{\text{AIV},k}$ of our AIV estimator depends on the corresponding true parameter value $\beta_{0,k}$. The red line in Figure 1 plots this relationship for one particular data generating process (DGP) that we also employ in our Monte Carlo simulations (the binary choice probit model with a continuous endogenous regressor of Table 1).[4] The details of this DGP do not matter here. What we are interested in are a couple of qualitative features of Figure 1 that are valid more generally:

[4]In that DGP both the variance of the endogenous regressor $X_{i,k}$ and of the error term U_i are equal to one.

(i) If the true value $\beta_{0,k}$ of the regression coefficient corresponding to the single endogenous regressor is equal to zero, then $\operatorname{plim} \widehat{\beta}_{\mathrm{AIV},k}$ is also equal to zero. We already know that this is true for all data generating processes that satisfy the conditions in Theorem 1.

(ii) According to Theorem 1, if the covariate $X_{i,k}$ is treated as endogenous and instrumented for, but is actually exogenous in the data-generating process (i.e. X_i is independent of U_i), then we have $\operatorname{plim} \widehat{\beta}_{\mathrm{AIV},k} = \beta_{0,k}$, corresponding to the 45-degree line drawn in grey in Figure 1. Therefore, if the degree of endogeneity is small (relative to the strengths of the instrument), then we would expect only a small deviation from the 45-degree line. Conversely, if the degree of endogeneity is large, then we generally expect larger deviations from the 45-degree line.

(iii) In Figure 1 the sign of $\operatorname{plim} \widehat{\beta}_{\mathrm{AIV},k}$ is always equal to the sign of $\beta_{0,k}$. If this property holds, then we say that $\widehat{\beta}_{\mathrm{AIV}}$ is "globally sign consistent". In our simulations in Section 5 we always find global sign consistency for all DGPs that we explore, but we are not able to provide formal conditions under which global sign consistency holds in this paper (apart from exogeneity of $X_{i,k}$). Instead, in the following we want to discuss "local sign consistency", that is, sign consistency in a small neighborhood of $\beta_{0,k} = 0$.

(iv) Local sign consistency of the AIV estimator leads to a test of the null hypothesis $H_0 : \beta_{0,k} = 0$ that is consistent for alternatives in a neighborhood of H_0. Global sign consistency leads to general consistency of the same test. This is particularly useful in applications where a main concern is whether the effect of an endogenous "treatment" variable is zero.

Let $\beta_*(\beta_0)$ be the large n probability limit of $\widehat{\beta}_{\mathrm{AIV}}$. We say that the k'th component of the AIV estimator is *locally sign consistent* if there exists $\delta > 0$ such that

$$\operatorname{sign}\left(\beta_{*,k}(\beta_0)\right) = \operatorname{sign}\left(\beta_{0,k}\right),$$

for all β_0 with $|\beta_{0,k}| < \delta$. Under appropriate smoothness conditions, a sufficient condition for local sign consistency of $\widehat{\beta}_{\mathrm{AIV},k}$ is given by

$$\left.\frac{\partial \beta_{*,k}(\beta_0)}{\partial \beta_{0,k}}\right|_{\beta_{0,k}=0} > 0. \tag{7}$$

In the following we give two concrete examples where (7) holds. Notice, however, that (7) is not a necessary condition for local sign consistency of $\widehat{\beta}_{\text{AIV},k}$, because one could, for example, have $\frac{\partial \beta_{*,k}(\beta_0)}{\partial \beta_{0,k}} = 0$, at $\beta_{0,k} = 0$, and still achieve local sign consistency via $\frac{\partial^2 \beta_{*,k}(\beta_0)}{\partial^2 \beta_{0,k}} = 0$ and $\frac{\partial^3 \beta_{*,k}(\beta_0)}{\partial^3 \beta_{0,k}} > 0$, at $\beta_{0,k} = 0$.

Example 1 (Probit control function model)**.** *Consider the generalized probit control function model:*

$$\begin{aligned} &Y_i = \mathbb{1}(X_i'\beta_0 - U_i > 0), \\ &x_i = g(Z_i, V_i), \quad X_i = (1, x_i), \\ &(U_i, V_i) \,\big|\, Z_i \sim (U_i, V_i) \sim F_{U,V},\ U_i \sim \mathcal{N}(0,1), \end{aligned} \tag{8}$$

where x_i, U_i, V_i are all scalar random variables, Z_i is a vector of instruments that includes a constant, g is strictly monotone in V_i, and $F_{U,V}$ is absolutely continuous with density $f_{U,V}$. This model is more general than the one studied in Rivers and Vuong (1988) in that it does not require the conditional distribution $U_i|V_i$ to be linear in V_i nor normal; the first-stage is allowed to be non-separable and non-linear in (Z_i, V_i), as in Imbens and Newey (2009) . In this example, the regressor $X_{i,k}$ for $k = 2$ is endogenous, and one can show (see Appendix) that

$$\left.\frac{\partial \beta_{*,2}(\beta_0)}{\partial \beta_{0,2}}\right|_{\beta_{0,2}=0} = 1,$$

therefore local sign consistency holds.

Example 2 (Generalized bivariate probit IV)**.** *Consider the bivariate probit IV model:*

$$\begin{aligned} &Y_i = \mathbb{1}(X_i'\beta_0 + U_i > 0), \\ &x_i = \mathbb{1}(m(Z_i) + V_i > 0) \quad X_i = (1, x_i),\ Z_i = (1, z_i), \\ &(U_i, V_i) \,\big|\, Z_i \sim (U_i, V_i) \sim F_{U,V},\ U_i \sim \mathcal{N}(0,1), \end{aligned} \tag{9}$$

where x_i, z_i, U_i, V_i are all scalar random variables, and $m(Z_i)$ is assumed to be a monotonic function of z_i. This model nests the popular bivariate probit model which further assumes joint normality of (U_i, V_i) and linearity of $m(Z_i)$. Again, the regressor $X_{i,k}$ for $k = 2$ is endogenous, and we show in the Appendix that (7) holds for $k = 2$, that is, local sign consistency holds in this example as well. Unlike Example 1, the arguments we use to show local sign consistency in this model do not directly generalize to the over-identified case ($k_z > 2$).

We know that local sign consistency of the AIV estimator holds whenever all the regressors are exogenous. In addition, the above examples provide two concrete data generating processses where a single regressor is endogenous and local sign consistency still holds. We have also verified local sign consistency (in fact, global sign consistency) numerically for all the data generating processes in our Monte Carlo simulations. We therefore conclude that local sign consistency of the AIV estimator holds for a large class of data generating processes.

As mentioned above, an important implication of the local sign consistency property is that a t-test for the hypothesis $H_0 : \beta_{0,k} = 0$ based on our estimator has non-trivial power — and it is in fact consistent — in a neighbourhood of the null hypothesis. The distribution of this t-test under H_0 is guaranteed by Theorem 2. For implementation, one just needs to compute the sample analog $\widehat{\text{AVar}}(\sqrt{n}\beta_{\text{AIV}})$ of the asymptotic variance-covariance matrix $(G'\,W\,G)^{-1}\,G'\,W\,\Sigma\,W\,G\,(G'\,W\,G)^{-1}$ given in the theorem, and $n(\widehat{\beta}_{\text{AIV},k})^2/[\widehat{\text{AVar}}(\sqrt{n}\beta_{\text{AIV}})]_{kk}$ will be $\chi^2(1)$ distributes as $n \to \infty$.

4 Generalization and implementation

We now want to discuss a generalization of the model and AIV estimator described in Section 2.1. Specifically, we now assume that in addition to (Y_i, X_i, Z_i), $i = 1, \ldots, n$, we also observe the additional strictly exogenous covariate W_i. The difference between X_i and W_i is that W_i need not enter the model through the linear single index $\omega_i = X_i'\beta$. Similarly, in addition to the unknown parameters β we now allow for the additional unknown parameters α, which also need not enter the model through the single index ω_i. Examples where this generalization is important are ordered choice models, Tobit models, and negative binomial models. The appropriate generalization of Assumption 1 is as follows:

Assumption 4 (Generalized Model).

(i) The outcomes Y_i are generated from the latent variable model

$$Y_i = g(\omega_{0,i}, W_i, U_i, \alpha_0), \qquad \omega_{i,0} := X_i'\,\beta_0,$$

where $U_i \in \mathbb{R}$ are unobserved random variables, the function $g(\cdot,\cdot,\cdot,\cdot)$ is known, and α_0 and β_0 are vectors of unknown parameters.

(ii) The distribution of U_i is independent of (Z_i, W_i), and U_i has known cdf $F_U(\cdot)$.

(iii) (X_i, Z_i, W_i, U_i) are independent and identically distributed across $i = 1, \ldots, n$.

Let $\ell\left(Y_i \,\middle|\, W_i,\, \omega_i,\, \alpha\right)$ be the log-likelihood of Y_i conditional on W_i, $\omega_{0,i} = \omega_i$ and $\alpha_0 = \alpha$. Then, the generalization of the AIV estimator in (1) is given by

$$\left(\widehat{\gamma}(\beta), \widehat{\alpha}(\beta)\right) = \underset{\gamma\in\mathcal{C},\alpha\in\mathcal{A}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta + Z_i'\gamma,\, \alpha\right), \quad \widehat{\beta}_{\mathrm{AIV}} \in \underset{\beta\in\mathcal{B}}{\operatorname{argmin}} \left\|\widehat{\gamma}(\beta)\right\|_{\Omega_{n,\beta}},$$
$$\widehat{\alpha}^{\dagger}(\beta) = \underset{\alpha\in\mathcal{A}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta,\, \alpha\right), \quad \widehat{\alpha}_{\mathrm{AIV}} = \widehat{\alpha}^{\dagger}(\widehat{\beta}_{\mathrm{AIV}}, 0), \tag{10}$$

where $\mathcal{B}$ and $\mathcal{C}$ are compact parameter sets as before and $\mathcal{A}$ is a compact parameter set for α. Compactness of the parameter sets is again a very helpful technical regularity condition to derive asymptotic results. However, for practical implementation we again assume that the boundedness imposed by $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{A}$ is not binding, that is, in practice we replace $\mathcal{B}$ by $\mathbb{R}^{k_x}$, $\mathcal{C}$ by $\mathbb{R}^{k_z}$ and $\mathcal{A}$ by $\mathbb{R}^{k_\alpha}$, where k_α denotes the dimension of α.

The appropriate generalizations of our consistency result of Theorem 1 for the AIV estimator in Section 3.1 to the model and estimator in Assumption 4 and display (10) are provided in the appendix.

In our Monte Carlo simulations and empirical applications below we focus on the binary choice model for which this extension of the model discussed here is not actually required. However, even for the binary choice model there can be computational advantages in implementing the AIV estimator according to (10) instead of (1). This is because we can move all the regression coefficients that correspond to exogenous covariates from β to α and then implement (10) instead of (1). The advantage of that implementation is that the inner-loop optimization over (γ, α) in (10) is a convex optimization problem (since we assume the log-likelihood to be a concave function) while the outer-loop optimization over β is in general a non-convex problem, implying that we want the dimension of the vector β to be as small as possible for computational reasons.

This computational issue is important in practice and we therefore want to be explicit about it. Consider the setup of our original Assumption 1 and decompose $X_i = (X_i^{\mathrm{end}\prime}, X_i^{\mathrm{ex}\prime})'$ and $Z_i = (Z_i^{\mathrm{ex}\prime}, X_i^{\mathrm{ex}\prime})$, where X_i^{end} are the endogenous regressors, X_i^{ex} are the exogenous regressors, and Z_i^{ex} are the excluded instruments. In most applications we expect X_i^{end} to be low-dimensional (often just a single variable). Let β^{end} and β^{ex} be the regression coefficients corresponding to X_i^{end} and X_i^{ex}. By applying the generalized AIV

estimator in (10) to this setup with $(X_i, W_i, Z_i, \beta, \alpha)$ equal to $(X_i^{\text{end}}, X_i^{\text{ex}}, Z_i^{\text{ex}}, \beta^{\text{end}}, \beta^{\text{ex}})$ we obtain

$$\begin{aligned}\left(\widehat{\gamma}(\beta^{\text{end}}), \overline{\beta}^{\text{ex}}(\beta^{\text{end}})\right) &= \underset{(\gamma, \beta^{\text{ex}})}{\text{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, X_i^{\text{end}\prime}\beta^{\text{end}} + X_i^{\text{ex}\prime}\beta^{\text{ex}} + Z_i^{\text{ex}\prime}\gamma\right), \\ \widehat{\beta}^{\text{end}} &\in \underset{\beta^{\text{end}}}{\text{argmin}} \left\|\widehat{\gamma}(\beta^{\text{end}})\right\|_{\Omega_{n,\beta}^{\text{end}}}, \\ \widehat{\beta}^{\text{ex}} &= \underset{\beta^{\text{ex}}}{\text{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, X_i^{\text{end}\prime}\widehat{\beta}^{\text{end}} + X_i^{\text{ex}\prime}\beta^{\text{ex}}\right), \end{aligned} \tag{11}$$

where $\Omega_{n,\beta}^{\text{end}}$ now is a positive definite matrix of dimension $\dim(X_i^{\text{end}}) \times \dim(X_i^{\text{end}})$ only.

Again, the key observation here is that the optimization over $(\gamma, \beta^{\text{ex}})$ is a convex optimization problem, while the optimization over β^{end} is non-convex but usually low-dimensional (often just one-dimensional which can e.g. be implemented by an initial grid-search followed by, for example, a golden-section search). Implementing the AIV estimator via (11) is therefore often computationally preferable to (1) and to (3), in particular, if k_x is large. Our results in the Appendix show that the two implementations are asymptotically equivalent when $k_z = k_x$. When $k_z > k_x$, then the choice of implementation and weight matrix matters for the (asymptotic) distribution of the resulting estimator, see the Appendix for more details.[5] In practice, we again recommend the simple choice $\Omega_{n,\beta}^{\text{end}} = \frac{1}{n}\sum_{i=1}^{n} Z_i^{\text{ex}} Z_i^{\text{ex}\prime}$.

5 Monte Carlo simulations

We consider the following data generating process (DGP):

$$\begin{aligned} &Y_i = \mathbb{1}\left\{\beta_1 + X_{2,i}\beta_2 + X_{3,i}\beta_3 + U_i \geq 0\right\}, \qquad U_i \sim \mathcal{N}(0,1), \\ &X_{2,i} = \sigma_{X_2}^{-1}\left(Z_i + V_i\right), \quad Z_i \sim (\chi^2(k) - k)/\sqrt{2k}, \quad k = 10, \quad X_{3,i} = \sigma_{X_3}^{-1}\left(\mathcal{N}(0,1) + 0.5 \cdot Z_i^2\right) \\ &V_i = \varepsilon_i + \delta_{end} \cdot (U_i + \delta_{no\_norm} \cdot (2 \cdot \mathbb{1}\left\{U_i \geq 0\right\} + U_i^2 - 2)), \quad \varepsilon_i \sim \mathcal{N}(0,1), \end{aligned}$$

with normalizing constants σ_{X_2} and σ_{X_3} chosen so that $\text{Var}(X_{2,i}) = \text{Var}(X_{3,i}) = 1$.

[5] When $k_z > k_x$, the asymptotic distribution for $\widehat{\beta}^{\text{end}}$ is equivalent under the two implementations if

$$\Omega = \begin{bmatrix} \Omega^{\text{end}} & 0 \\ 0 & \Omega^{\text{ex}} \end{bmatrix} \quad \text{and} \quad \Sigma_{\gamma\alpha} := \mathbb{E}\left[Z_i X_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i^{\text{end}\prime}\beta_0^{\text{end}} + X_i^{\text{ex}\prime}\beta_0^{\text{ex}}\right)}{\partial \omega^2}\right] = 0.$$

Beyond this set of special conditions, the two implementations do not in general lead to asymptotically equivalent estimators under over-identification.

We set $\beta_1 = 1$, $\beta_3 = -1$ and we document the performance of different procedures in the estimation of β_2 under different configurations of β_2, δ_{end} and δ_{no_norm}. We also report the empirical size of a two-sided t-test for the null hypothesis that β_2 is equal to its true value.

The AIV estimator is implemented as in (11), with outer-loop direct search over β_2 initialized at the control function estimate. Standard errors used in the t-test are based on the sample analogue of the asymptotic variance formula in Theorem 2.

For the control function estimator, the test statistic is based on the standard error formula provided in Rivers and Vuong (1988), which assumes correct specification of the model (including joint normality).[6]

The results are collected in Table 1. As expected, MLE is severely biased under endogeneity of the regressor and non-normality of the errors, leading to confidence intervals with no coverage. As predicted by theory, the control function estimator is consistent and provides accurate inference under joint normality of the errors, or in the absence of endogeneity. However, the coverage of its associated confidence intervals is null in the presence of endogeneity and lack of joint normality, due to large biases. The AIV estimator instead enjoys negligible bias under all configurations considered, at the cost of mild variance increases compared to the control function approach. Remarkably, the resulting rejection probabilities for a two-sided t-test are close to nominal size, including for values of β_2 away from 0. Figure 2 reports the power function of a two-sided t-test of regressor relevance ($H_0 : \beta_2 = 0$) based on the AIV estimator under $\delta_{end} = 1$ and $\delta_{no\_norm} = 2$. The sign-consistency property of the AIV estimator results in good power for this test, even though the presence of bias in the estimator for values of β_2 away from 0 leads to non-monotonic power in this DGP.

[6]Notice that the asymptotic variance formula contained in Rivers and Vuong (1988) is for a different normalization of the variance of U_i compared to MLE, bivariate Probit and the AIV estimators, which all assume $\text{Var}(U_i) = 1$. In order to make the control function estimates comparable with the other methods, we rescale the original control function estimates based on the normalization of Rivers and Vuong (1988) and appropriately adjust standard errors via the Delta method.

5.1 Simulations with binary endogenous regressor

We consider a modification of the previous DGP which now features a binary endogenous regressor:

$$Y_i = \mathbb{1}\left\{\beta_1 + X_{2,i}\beta_2 + X_{3,i}\beta_3 + U_i \geq 0\right\}, \qquad U_i \sim \mathcal{N}(0,1),$$
$$X_{2,i} = \left\{\sigma_{X_2}^{-1}\left(Z_i + V_i\right) \geq 0\right\}, \quad Z_i \sim (\chi^2(k) - k)/\sqrt{2k}, \quad k = 10, \quad X_{3,i} = \sigma_{X_3}^{-1}\left(\mathcal{N}(0,1) + 0.5 \cdot Z_i^2\right)$$
$$V_i = \varepsilon_i + \delta_{end} \cdot (U_i + \delta_{no\_norm} \cdot (2 \cdot \mathbb{1}\left\{U_i \geq 0\right\} + U_i^2 - 2)), \quad \varepsilon_i \sim \mathcal{N}(0,1).$$

where $\sigma_{X_2}, \sigma_{X_3}, \beta_3$ are as before, and we set $\beta_1 = 0.4$ to ensure $\mathbb{E}[Y_i] \approx 0.5$.

The results are given in Table 2. Surprisingly, the control function estimator has negligible bias under endogeneity and non-normality.[7] However, the associated rejection probabilities for the control function estimator are far from nominal size due to severe underestimation of the standard errors. As expected, the bivariate probit estimator performs well under joint-normality of the errors or exogeneity of the regressors. Under endogeneity and non-normality, the bivariate probit estimator suffers from large bias and considerable size distortions of its associated tests. On the other hand, the AIV estimator has negligible bias, resulting in good size control and high power of the associated two-sided test of regressor relevance, as shown in Figure 3. It is interesting to notice that 2SLS is not sign-consistent for the effect of the endogenous treatment in this DGP, whilst it is known to be sign-consistent in the absence of additional covariates (Bhattacharya, Shaikh and Vytlacil, 2012). The AIV estimator enjoys sign-consistency in this DGP, suggesting improved robustness of the sign-consistency property to the inclusion of additional covariates compared to 2SLS.

6 Empirical applications

In this section we present two empirical applications. In each application we compare estimates of the coefficient on the the binary endogenous regressor of interest based on popular existing estimators and the AIV estimator. In the first application, a test of relevance of the endogenous regressor based on the AIV estimator cautions the researcher about the conclusion that having health insurance increases the probability that an individual visits a doctor in a given year. In the second application, the AIV estimator

[7] We have verified that this small bias property exhibited by the control function estimator in this DGP is coincidental, and does not hold generally.

confirms the conclusion that smoking habits are transmitted by a mother to her offspring, a conclusion that would also be reached using existing methods. Overall, the two applications showcase the usefulness of the AIV estimator as a tool for checking the robustness of inferential conclusions in nonlinear models.

6.1 The effect of health insurance on hospital visits (Han and Lee, 2019)

Health insurance coverage is considered an important factor for patients' decisions to use medical services. On the other hand, the decision to acquire health insurance is endogenously determined by an individual's health status, as well as socioeconomic characteristics that are correlated with health outcomes. In this application, we investigate how health insurance coverage affects an individual's choice to visit a doctor. For this purpose, we use a dataset constructed by Han and Lee (2019) which combines data from the 2010 wave of the Medical Expenditure Panel Survey (MEPS) with information from the National Compensation Survey published by the US Bureau of Labor Statistics. The outcome of interest Y_i is a binary variable indicating whether an individual visited a doctor's office in January 2010; the binary endogenous treatment X_i^{end} indicates whether an individual has his/her own private insurance. Two instrumental variables are used following Zimmer (2018): the number of employees in the firm at which the individual works and a dummy variable that indicates whether a firm has multiple locations. These variables reflect how big the firm is, and the underlying rationale for using these variables as instruments is a that a bigger the firm is more likely it provides fringe benefits including health insurance. The validity of these instruments relies on firm size not directly affecting the decision to visit a doctor. Following Han and Lee (2019), we include a further 23 exogenous variables in the model as additional controls, including demographic characteristics as well as indicators of health status.

Table 3 provides estimates for the coefficient β^{end} on the binary treatment using probit MLE, 2SLS, the control function estimator of Rivers and Vuong (1988), the bivariate probit estimator and the AIV estimator. We also report the associated standard errors and the p-value of a two-sided t-test of no effect of health insurance coverage on doctor visits ($H_0 : \beta^{\text{end}} = 0$). Remarkably, all methods deliver positive estimates for β^{end} with similar magnitudes, with the exception of MLE being roughly three times smaller than the

other methods considered.[8] The test of regressor relevance based on bivariate probit leads to rejection of the null hypothesis at all conventional levels of significance. On the other hand, a test based on the AIV estimator does not reject the same hypothesis at the 1% level of significance. The difference between p-values in this application is driven by the varying magnitude of the standard errors associated with each method. Standard errors associated with bivariate probit are likely to underestimate the sampling variability of the estimator, as their validity relies on the assumptions of joint normality of the unobserved disturbances and linearity of the first-stage equation. Our theory reassures us that the AIV estimator provides inference that is robust to relaxing those assumptions in this application.

6.2 The intergenerational transmission of smoking habits (Mu and Zhang, 2018)

Vertical transmission within family is considered a key driver of the persistence of health behaviours. The way in which harmful practices such as smoking are transmitted within a family has therefore important implications for health policies. In this application we apply our proposed methods to the study of the intergenerational transmission of smoking habits using data from British Household Panel Survey. The outcome of interest Y_i is a binary variable indicating whether an adolescent smokes or not; the binary endogenous treatment X_i^{end} indicates whether his/her single mother smokes or not. Following Loureiro, Sanz-de Galdeano and Vuri (2010) and Mu and Zhang (2018), the instrument used is an indicator for whether the teenagers' grandfather had a high-skilled or low-skilled occupation (including unemployed). The underlying rationale for using this variable as an instrument is that the impact of parental socio-economic status on smoking behaviour does not extend beyond one generation, after controlling for the relevant explanatory variables. We include a further 5 exogenous variables in the model as additional controls: the child's age at interview year, the single mother's age at interview year, an indicator for whether the mother has higher education, an indicator for whether the mother is in a high-skilled or low-skilled occupation, and the natural logarithm of monthly household

[8] As an estimator for the average partial effect of X^{end} rather than the coefficient β^{end}, only the sign of the 2SLS estimator can be compared to the other estimators. Even though we report results for the control function estimator, its use is not recommended in this application as the endogenous regressor is binary.

income. Table 4 provides estimation results for the coefficient β^{end}. All methods deliver positive estimates for the coefficient β^{end}, implying that a mother's decision to smoke increases the probability that her offspring chooses to be a smoker too. Similarly to the previous empirical application, we find that all methods deliver estimates of similar magnitude, with the exception of the MLE estimate being roughly a third of the bivariate probit and AIV estimators. While the AIV estimator delivers a smaller estimate for β^{end} compared to bivariate probit, two-sided tests based on these two estimators both lead to rejection of the hypothesis $H_0 : \beta^{\text{end}} = 0$ at all conventional levels of significance. As a result, the AIV estimator provides evidence on the robustness of the conclusion that smoking habits are transmitted between generations.

7 Conclusions

We have introduced the AIV estimator as a new and simple estimator in non-linear models with endogenous covariates. The estimator translates the concept of excluded instruments into a criterion function that demands the MLE of the coefficients of the instruments to be close to zero when the instruments are included as covariates. We show that the resulting AIV estimator is consistent if the endogenous regression coefficients are equal to zero. For the case of a single endogenous regressor, we also demonstrate that the AIV estimator is usually sign-consistent. These properties and its simplicity make the estimator useful in practice, as illustrated by our empirical applications. In particular, the estimator is complementary to the control function and the probit IV estimator, because it makes weaker assumptions, but also delivers weaker consistency results.

References

Abrevaya, J., J. A. Hausman, and S. Khan (2010). Testing for causal effects in a generalized regression model with endogenous regressors. *Econometrica 78*(6), 2043–2061.

Angrist, J. D. and J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives 24*(2), 3–30.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2012). Treatment effect bounds: An application to swan–ganz catheterization. *Journal of Econometrics 168*(2), 223–243.

Chernozhukov, V. and C. Hansen (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics 132*(2), 491–525.

Chesher, A. (2010). Instrumental variable models for discrete outcomes. *Econometrica 78*(2), 575–601.

Chesher, A. and A. M. Rosen (2017). Generalized instrumental variable models. *Econometrica 85*(3), 959–989.

Dai, J. Y. and X. C. Zhang (2015). Mendelian randomization studies for a continuous exposure under case-control sampling. *American Journal of Epidemiology 181*(6), 440–449.

Han, S. and S. Lee (2019). Estimation in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Applied Econometrics 34*(6), 994–1015.

Han, S. and E. J. Vytlacil (2017). Identification in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Econometrics 199*(1), 63–73.

Harding, M. and C. Lamarche (2014). Estimating and testing a quantile regression model with interactive effects. *Journal of Econometrics 178*, 101–113.

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

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Lee, N., H. R. Moon, and M. Weidner (2012). Analysis of interactive fixed effects dynamic linear panel regression with measurement error. *Economics Letters 117*(1), 239–242.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics 97*(1), 145–177.

Little, R. J. A. (1985). A note about models for selectivity bias. *Econometrica 53*(6), 1469–1474.

Loureiro, M. L., A. Sanz-de Galdeano, and D. Vuri (2010). Smoking habits: Like father, like son, like mother, like daughter?*. *Oxford Bulletin of Economics and Statistics 72*(6), 717–743.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics 27*(3), 313–333.

Monfardini, C. and R. Radice (2008). Testing exogeneity in the bivariate probit model: A monte carlo study*. *Oxford Bulletin of Economics and Statistics 70*(2), 271–282.

Moon, H. R., M. Shum, and M. Weidner (2018). Estimation of random coefficients logit demand models with interactive fixed effects. *Journal of Econometrics 206*(2), 613–644.

Mourifié, I. and R. Méango (2014). A note on the identification in two equations probit model with dummy endogenous regressor. *Economics Letters 125*(3), 360–363.

Mu, B. and Z. Zhang (2018, 06). Identification and estimation of heteroscedastic binary choice models with endogenous dummy regressors. *The Econometrics Journal 21*(2), 218–246.

Newey, W. K. (1986). Linear instrumental variable estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics 32*(1), 127–141.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Rivers, D. and Q. H. Vuong (1988). Limited information estimators and exogeneity tests for simulataneous probit models. *Journal of Econometrics 39*(1), 347–366.

Windmeijer, F. (2019). Two-stage least squares as minimum distance. *The Econometrics Journal 22*(1), 1–9.

Wright, P. G. (1928). *Tariff on animal and vegetable oils*. Macmillan Company, New York.

Yildiz, N. (2013). Estimation of binary choice models with linear index and dummy endogenous variables. *Econometric Theory 29*(2), 354–392.

Zimmer, D. (2018). Using copulas to estimate the coefficient of a binary endogenous regressor in a poisson regression: Application to the effect of insurance on doctor visits. *Health Economics 27*(3), 545–556.

Table 1: Monte Carlo simulations with continuous endogenous regressor, $n = 7000$

			MLE			2SLS			Control Function			Auxiliary IV		
β_2	δ_{end}	$\delta_{\substack{no \\ norm}}$	Bias	Std.	$\hat{p}; .05$	Bias	Std.	$\hat{p}; .05$	Bias	Std.	$\hat{p}; .05$	Bias	Std.	$\hat{p}; .05$
0	1	0	−0.74	.024	1.00	.016	.009	.415	.000	.031	.047	.000	.033	.047
0	1	2	−1.22	.054	1.00	.041	.024	.385	.550	.030	.999	−.002	.079	.046
−0.1	1	2	−1.22	.063	1.00	.106	.024	.999	.499	.025	1.00	−.006	.077	.047
0.1	1	2	−1.19	.046	1.00	.022	.025	.168	.590	.058	.99	.031	.087	.059
1	0	-	.000	.025	.046	.756	.006	1.00	.000	.032	.049	.001	.043	.047

Notes: Simulation results based on 5000 replications.

Table 2: Monte Carlo simulations with binary endogenous regressor, $n = 7000$

			2SLS			Control Function			Bivariate Probit			Auxiliary IV		
β_2	δ_{end}	$\delta_{\substack{no \\ norm}}$	Bias	Std.	$\hat{p}; .05$	Bias	Std.	$\hat{p}; .05$	Bias	Std.	$\hat{p}; .05$	Bias	Std.	$\hat{p}; .05$
0	1	0	.124	.025	.990	.152	.076	.521	.001	.064	.054	.001	.082	.052
0	1	2	.310	.069	.999	−.020	.399	.513	.325	.178	.494	.006	.209	.048
−0.1	1	2	.370	.068	1.00	−.053	.399	.440	.403	.183	.748	−.029	.209	.054
0.1	1	2	.240	.071	.973	−.013	.412	.521	.281	.121	.170	.046	.216	.042
1	0	-	.690	.039	1.00	.000	.069	.056	.000	.065	.057	.001	.072	.056

Notes: Simulation results based on 5000 replications.

Table 3: Effect of Health Insurance on Doctor Visits

	$\widehat{\beta}^{\text{end}}$	Std. Err.	p-value
MLE	.1796	.0404	$< .0000$
2SLS	.1326	.0459	.0038
Control Function	.5358	.1740	.0020
Bivariate Probit	.4962	.1558	.0014
Auxiliary IV	.5622	.2487	.0238

Sample size $n = 7555$. Data source: Han and Lee (2019).

Table 4: Effect of mother's smoking habits on child's smoking habits

	$\widehat{\beta}^{\text{end}}$	Std. Err.	p-value
MLE	.3305	.0347	$< .0000$
2SLS	.3746	.1243	.0026
Control Function	1.089	.3047	.0004
Bivariate Probit	1.440	.1203	$< .0000$
Auxiliary IV	1.130	.4269	.0081

Sample size $n = 7053$. Data source: Mu and Zhang (2018).

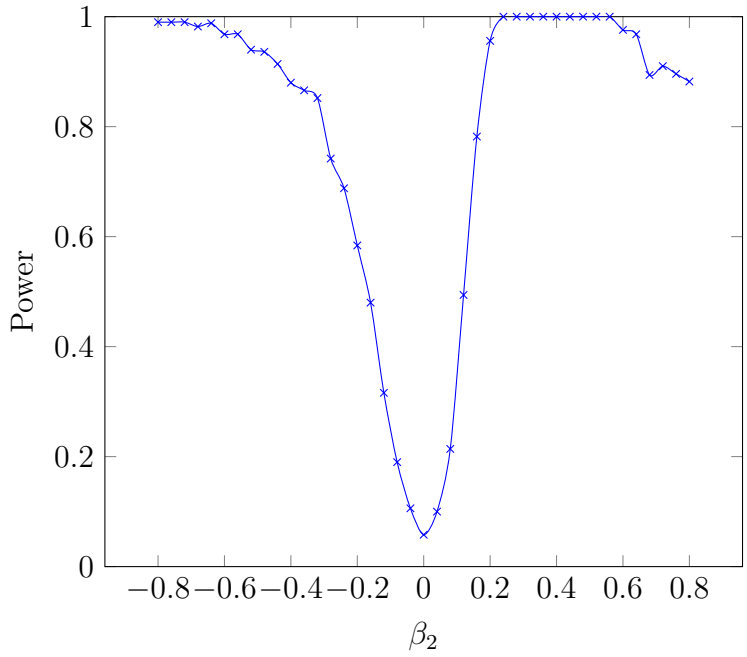

Figure 2: Power function of two-sided test with continuous endogenous regressor, $n = 7000$

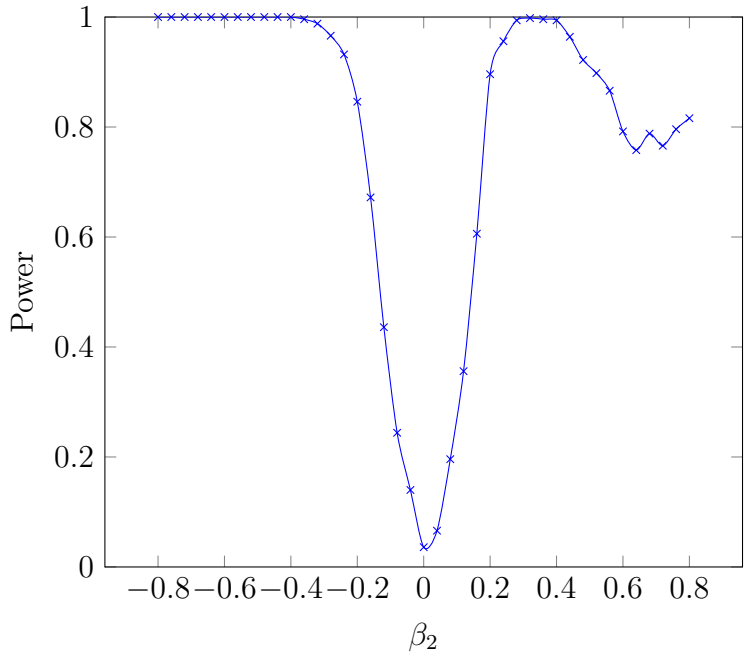

Figure 3: Power function of two-sided test with discrete endogenous regressor, $n = 7000$

A Appendix

A.1 Consistency result for generalized model

Here we present a consistency result for the generalized model of Assumption 4. We make the following assumptions.

Assumption 5 (Exogeneity of $X_i'\beta_0$). *U_i is independent of $(W_i, X_i'\beta_0, Z_i)$.*

Assumption 6 (Regularity conditions).

(i) The parameter sets $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{A}$ are compact. $\mathcal{B}$ contains β_0, $\mathcal{C}$ contains 0 and $\mathcal{A}$ contains α_0 as interior points respectively.

(ii) For all possible outcomes y, the log-likelihood function $\ell\left(y \,\middle|\, w, \omega, \alpha\right)$ is strictly convex in (ω, α) and has Hessian with eigenvalues bounded away from zero, uniformly over (w, ω, α). Furthermore, $\ell\left(Y_i \,\middle|\, W_i, X_i'\beta + Z_i'\gamma, \alpha\right)$ is three times continuously differentiable in (β, γ, α) with derivatives that in expectation are bounded for all $(\beta, \gamma, \alpha) \in \mathcal{B} \times \mathcal{C} \times \mathcal{A}$.

(iii) Let $\mathcal{L}(\beta, \gamma, \alpha) := \mathbb{E}\left[\ell\left(Y_i \,\middle|\, W_i, X_i\beta + Z_i\gamma,\, \alpha\right)\right]$ denote the population log-likelihood function. For all $\beta \in \mathcal{B}$ and $\eta := (\gamma, \alpha) \in \mathcal{C} \times \mathcal{A}$, we have

$$\operatorname{rank}\left\{\frac{\partial^2 \mathcal{L}(\beta, \gamma, \alpha)}{\partial \eta \partial \eta'}\right\} = k_z + k_\alpha. \tag{12}$$

For all $\beta \in \mathcal{B}$ and $(0, \alpha) \in \mathcal{C} \times \mathcal{A}$, the matrix

$$A(\beta, 0, \alpha) := \begin{pmatrix} \frac{\partial^2 \mathcal{L}(\beta,0,\alpha)}{\partial \gamma \partial \beta'} & \frac{\partial^2 \mathcal{L}(\beta,0,\alpha)}{\partial \gamma \partial \alpha'} \\ \frac{\partial^2 \mathcal{L}(\beta,0,\alpha)}{\partial \alpha \partial \beta'} & \frac{\partial^2 \mathcal{L}(\beta,0,\alpha)}{\partial \alpha \partial \alpha'} \end{pmatrix} \tag{13}$$

has full rank $k_x + k_\alpha$.

(iv) The second derivatives of the sample log-likelihood

$$\mathcal{L}_n(\beta, \gamma, \alpha) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha\right)$$

converge in probability to those of the population log-likelihood

$$\mathcal{L}(\beta, \gamma, \alpha) = \mathbb{E}\left[\ell\left(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha\right)\right]$$

uniformly over $(\beta, \gamma, \alpha) \in \mathcal{B} \times \mathcal{C} \times \mathcal{A}$.

(v) The symmetric matrix $\Omega_{n,\beta}$ is a twice continuously differentiable function in β, and there exists a constant $c > 0$ such that with probability approaching one we have $\Omega_{n,\beta} \geq c$ for all $\beta \in \mathcal{B}$. Furthermore, we have $\sup_{\beta\in\mathcal{B}} \|\Omega_{n,\beta} - \Omega_\beta\| = o_p(1)$ for some non-random symmetric matrix Ω_β which is positive-definite for all $\beta \in \mathcal{B}$.

Assumptions 5 and 6 generalize Assumptions 2 and 3, respectively, to the case with additional regressors W_i and parameters α. In particular, Assumption 6(iii) imposes generalizations of the non-collinearity and relevance conditions for the instruments. Under (12), condition (13) is equivalent to requiring

$$\operatorname{rank}\left\{\frac{\partial^2\mathcal{L}(\beta,0,\alpha)}{\partial\gamma\partial\beta'} - \left[\frac{\partial^2\mathcal{L}(\beta,0,\alpha)}{\partial\gamma\partial\alpha'}\right]\left[\frac{\partial^2\mathcal{L}(\beta,0,\alpha)}{\partial\alpha\partial\alpha'}\right]^{-1}\left[\frac{\partial^2\mathcal{L}(\beta,0,\alpha)}{\partial\alpha\partial\beta'}\right]\right\} = k_x. \tag{14}$$

Theorem 3. *Let Assumption 4, 5, 6 hold. Then we have $(\widehat{\beta}_{\text{AIV}}, \widehat{\alpha}_{\text{AIV}}) = (\beta_0, \alpha_0) + o_P(1)$, as $n \to \infty$.*

A.2 Asymptotic normality result for generalized model

We now present the general result for the asymptotic distribution of the AIV estimator. To do so, we introduce the following notation for the first and second derivatives of the sample and population log-likelihood:

$$\mathcal{L}_\alpha(\beta,\gamma,\alpha) = \mathbb{E}\left[\frac{\partial\ell(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha)}{\partial\alpha}\right] \qquad \mathcal{L}_{n,\alpha}(\beta,\gamma,\alpha) = \frac{1}{n}\sum_{i=1}^n \frac{\partial\ell(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha)}{\partial\alpha},$$

$$\mathcal{L}_{\alpha\beta}(\beta,\gamma,\alpha) = \mathbb{E}\left[\frac{\partial\ell(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha)}{\partial\alpha\,\partial\beta'}\right] \qquad \mathcal{L}_{n,\alpha\beta}(\beta,\gamma,\alpha) = \frac{1}{n}\sum_{i=1}^n \frac{\partial\ell(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha)}{\partial\alpha\,\partial\beta'},$$

where we will also use the short-hand $\mathcal{L}_{\alpha\beta} \coloneqq \mathcal{L}_{\alpha\beta}(\beta_0, 0, \alpha_0)$. We also define the matrices

$$\widetilde{H} = \mathcal{L}_{\gamma\gamma} - \mathcal{L}_{\gamma\alpha}\,\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\gamma}, \quad \widetilde{G} = \mathcal{L}_{\gamma\beta} - \mathcal{L}_{\gamma\alpha}\,\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta},$$

and their sample analogues $\widetilde{H}_n, \widetilde{G}_n$ based on $\mathcal{L}_{n,\alpha\alpha}, \mathcal{L}_{n,\alpha\beta}, \ldots$ in the natural way.

Theorem 4. *Let Assumption 4, 5, 6 hold. Then we have*

$$\begin{aligned}
\sqrt{n}\,(\widehat{\beta}_{\text{AIV}} - \beta_0) &= -(\widetilde{G}'\,\widetilde{W}\,\widetilde{G})^{-1}\,\widetilde{G}'\,\widetilde{W}\,\sqrt{n}\,\left\{\mathcal{L}_{n,\gamma} - \mathcal{L}_{\gamma\alpha}\,\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{n,\alpha}\right\} + o_p(1),\\
\sqrt{n}\,(\widehat{\alpha}_{\text{AIV}} - \alpha_0) &= -\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta}\,\sqrt{n}\,(\widehat{\beta}_{\text{AIV}} - \beta_0) - \mathcal{L}_{\alpha\alpha}^{-1}\,\sqrt{n}\mathcal{L}_{n,\alpha} + o_p(1)\\
&= \mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta}\,(\widetilde{G}'\,\widetilde{W}\,\widetilde{G})^{-1}\,\widetilde{G}'\,\widetilde{W}\,\sqrt{n}\,\mathcal{L}_{n,\gamma}\\
&\qquad - \left\{\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta}\,(\widetilde{G}'\,\widetilde{W}\,\widetilde{G})^{-1}\,\widetilde{G}'\,\widetilde{W}\,\mathcal{L}_{\gamma\alpha}\mathcal{L}_{\alpha\alpha}^{-1} + \mathcal{L}_{\alpha\alpha}^{-1}\right\}\,\sqrt{n}\,\mathcal{L}_{n,\alpha} + o_p(1),
\end{aligned}$$

where $\widetilde{W} := \widetilde{H}^{-1}\,\Omega_{\beta_0}\,\widetilde{H}^{-1}$.

The asymptotic representation in Theorem 4 can be used to show asymptotic normality of the AIV estimator based on

$$\sqrt{n}\begin{pmatrix}\mathcal{L}_{n,\gamma}\\ \mathcal{L}_{n,\alpha}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\Sigma_\gamma & \Sigma_{\gamma\alpha}\\ \Sigma_{\gamma\alpha}' & \Sigma_\alpha\end{pmatrix}\right]$$

where $\Sigma_\alpha = \text{Var}\left[\frac{\partial \ell(Y_i|W_i, X_i'\beta_0, \alpha_0)}{\partial \alpha}\right]$, $\Sigma_\gamma = \text{Var}\left[Z_i \frac{\partial \ell\left(Y_i \,|\, W_i, X_i'\beta_0, \alpha_0\right)}{\partial \omega}\right]$, and $\Sigma_{\gamma\alpha}$ was defined in the main text.

A.3 Local sign consistency: formal results

In this section, we formalize conditions under which the auxiliary IV estimator is sign-consistent and we show that these conditions are verified in two benchmark models. For this purpose we need some additional notation. Let $\gamma(\cdot,\cdot) : \mathbb{R}^{k_x} \times \mathbb{R}^{k_x} \to \mathbb{R}^{k_z}$ be the function implicitly defined by the relationship

$$s(\beta, \gamma(\beta,\beta_0), \beta_0) = 0, \quad s(\beta,\gamma,\beta_0) := \mathbb{E}_{P_{\beta_0}}\left[\frac{\partial \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega} Z_i\right],$$

where P_{β_0} denotes the true data generating process parametrized by β_0. Our previous results show that $\widehat{\gamma}(\beta)$ defined in (1) converges uniformly to $\gamma(\beta,\beta_0)$ under P_{β_0}. Thus, we can define the probability limit of our auxiliary IV estimator as a function of β_0 as

$$\beta^*(\beta_0) = \underset{\beta\in\mathcal{B}}{\text{argmin}} \left\|\gamma(\beta,\beta_0)\right\|_{\Omega(\beta,\beta_0)}, \tag{15}$$

where $\Omega(\beta,\beta_0)$ is the probability limit of $\Omega_{n,\beta}$ under P_{β_0}. The next theorem provides a sufficient condition for local sign consistency of the AIV estimator, which relies on some additional regularity conditions.

Assumption 7 (Additional regularity conditions). *There exists an open set around $\beta^{\mathcal{O}} = (\beta^{\mathcal{O}}_{-k}, 0)$ such that*

(i) The function $s(\beta,\gamma,\beta_0)$ is three-times continuously differentiable with uniformly bounded derivatives, and second derivatives

$$G(\beta,\gamma,\beta_0) := \frac{\partial^2 s(\beta,\gamma,\beta_0)}{\partial\gamma\partial\beta'} = \mathbb{E}_{P_{\beta_0}}\left[Z_i X_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i\beta + Z_i\gamma\right)}{\partial\omega^2}\right],$$

$$H(\beta,\gamma,\beta_0) := \frac{\partial^2 s(\beta,\gamma,\beta_0)}{\partial\gamma\partial\beta'} = \mathbb{E}_{P_{\beta_0}}\left[Z_i Z_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i\beta + Z_i\gamma\right)}{\partial\omega^2}\right],$$

having singular values uniformly bounded away from 0.

(ii) The function $\Omega(\beta, \beta_0)$ is positive-definite with eigenvalues uniformly bounded away from 0 and uniformly bounded entries that have uniformly bounded continuous derivatives up to second-order.

Assumption 7 imposes additional smoothness conditions on the population score function. These guarantee that the AIV estimator solves a convex optimization problem for data generating process in a neighborhood of $\beta^{\mathcal{O}}$, when the optimization is made over a suitably small set.

Theorem 5. *Suppose that Assumptions 1,2, 3, and 7 hold. Then the auxiliary IV estimator that solves (1) over a suitably small $\mathcal{B}_* \subseteq \mathcal{B}$ is locally sign consistent if*

$$\left.\frac{\partial \beta_{*,k}(\beta_0)}{\partial \beta_{0,k}}\right|_{\beta_0=\beta^{\mathcal{O}}} = \left[\left(G_{\mathcal{O}}' H_{\mathcal{O}}^{-1} \Omega_{\mathcal{O}} H_{\mathcal{O}}^{-1} G_{\mathcal{O}}\right)^{-1} G_{\mathcal{O}}' H_{\mathcal{O}}^{-1} \Omega_{\mathcal{O}} H_{\mathcal{O}}^{-1} \frac{\partial s(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})}{\partial \beta_0'}\right]_{(k_x, k_x)} > 0,$$

where $G_{\mathcal{O}} = G(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})$, $H_{\mathcal{O}} = H(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})$ and $\Omega_{\mathcal{O}} = \Omega(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})$.

In the following subsections, we use the above Lemma to verify local sign consistency of the auxiliary IV estimator in the benchmark models of Examples 1 and 2.

A.3.1 Details for Example 1 (Control function)

We have

$$\begin{aligned} s(\beta, \gamma, \beta_0) &= \mathbb{E}_{P_{\beta_0}}\left[(Y_i - \Phi(X_i'\beta + Z_i'\gamma)) \cdot \frac{\phi(X_i'\beta + Z_i'\gamma)}{\Phi(X_i'\beta + Z_i'\gamma) \cdot (1 - \Phi(X_i'\beta + Z_i'\gamma))} Z_i\right] \\ &= \mathbb{E}_{P_{\beta_0}}\left[\left\{F_{U|V}(X'\beta_0 \mid V_i) - \Phi(X_i'\beta + Z_i'\gamma)\right\} \cdot \frac{\phi(X_i'\beta + Z_i'\gamma)}{\Phi(X_i'\beta + Z_i'\gamma) \cdot (1 - \Phi(X_i'\beta + Z_i'\gamma))} Z_i\right], \end{aligned}$$

since

$$\begin{aligned} \mathbb{E}_{P_{\beta_0}}[Y_i \mid X_i, Z_i] &= F_{U|X,Z}(X'\beta_0 \mid X_i, Z_i) \\ &= F_{U|V,Z}(X'\beta_0 \mid V_i, Z_i) \\ &= F_{U|V}(X'\beta_0 \mid V_i), \end{aligned}$$

which then gives

$$\frac{\partial s(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})}{\partial \beta_0} = \frac{\phi(\beta_2^{\mathcal{O}})}{\Phi(\beta_2^{\mathcal{O}}) \cdot (1 - \Phi(\beta_2^{\mathcal{O}}))} \cdot \mathbb{E}\left[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, Z_i X_i'\right].$$

Having defined $\widetilde{Z}_i := \Omega_{\mathcal{O}}^{-1/2} H_{\mathcal{O}}^{-1} Z_i$, we have by Theorem 5 that

$$\left.\frac{d\beta_*(\beta_0)}{d\beta_0}\right|_{\beta_0=\beta_{\mathcal{O}}} = \frac{1}{\phi(\beta_2^{\mathcal{O}})} \cdot \left[\mathbb{E}[X_i \widetilde{Z}_i'] \cdot \mathbb{E}[\widetilde{Z}_i' X_i]\right]^{-1} \mathbb{E}[X_i \widetilde{Z}_i'] \cdot \mathbb{E}\left[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, \widetilde{Z}_i X_i'\right], \quad (16)$$

It is useful to define $Q := \left[\mathbb{E}[X_i \widetilde{Z}_i'] \cdot \mathbb{E}[\widetilde{Z}_i' X_i]\right]$, for which we we have

$$Q^{-1} = \frac{1}{\det(Q)} \cdot \begin{pmatrix} Q_{22} & -Q_{12} \\ -Q_{12} & Q_{11} \end{pmatrix},$$

$$Q_{11} = \sum_{j=1}^{k_z} \mathbb{E}[x_i\, \widetilde{Z}_{i,j}]^2, \quad Q_{12} = \sum_{j=1}^{k_z} \mathbb{E}[x_i\, \widetilde{Z}_{i,j}] \cdot \mathbb{E}[\widetilde{Z}_{i,j}], \quad Q_{22} = \sum_{j=1}^{k_z} \mathbb{E}[\widetilde{Z}_{i,j}]^2.$$

We also have

$$\begin{aligned}
\mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, Z_i x_i] &= \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, m(Z_i) \widetilde{Z}_i] + \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i \widetilde{Z}_i] \\
&= \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)] \cdot \mathbb{E}[m(Z_i) \widetilde{Z}_i] + \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i] \cdot \mathbb{E}[\widetilde{Z}_i] \\
&= \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)] \cdot \mathbb{E}[x_i \widetilde{Z}_i] + \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i] \cdot \mathbb{E}[\widetilde{Z}_i] \\
&= \phi(\beta_2^{\mathcal{O}}) \cdot \mathbb{E}[x_i \widetilde{Z}_i] + \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i] \cdot \mathbb{E}[\widetilde{Z}_i],
\end{aligned}$$

where we have used the independence between Z_i and V_i. Thus we can express the lower-diagonal entry of (16)

$$\left.\frac{\partial \beta_{*,2}(\beta_0)}{\partial \beta_{0,2}}\right|_{\beta_0=\beta_{\mathcal{O}}} = \left[Q^{-1}\mathbb{E}[X_i \widetilde{Z}_i']\right]_{(2,\bullet)} \cdot \mathbb{E}\left[\widetilde{Z}_i X_i\right]_{(\bullet,2)} + \frac{\mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i]}{\phi(\beta_2^{\mathcal{O}})} \cdot \left[Q^{-1}\mathbb{E}[X_i \widetilde{Z}_i']\right]_{(2,\bullet)} \cdot \mathbb{E}[\widetilde{Z}_i].$$

The first term in the above expansion is equal to 1 and the second term is equal to 0 since

$$\left[Q^{-1}\mathbb{E}[X_i \widetilde{Z}_i']\right]_{(2,\bullet)} \cdot \mathbb{E}[\widetilde{Z}_i] = \frac{1}{\det(Q)} \cdot \left[- Q_{12} \cdot \left(\sum_{j=1}^{k_z} \mathbb{E}[\widetilde{Z}_{i,j}]^2 \right) + Q_{11} \cdot \left(\sum_{j=1}^{k_z} \mathbb{E}[x_i \widetilde{Z}_{i,j}] \cdot \mathbb{E}[\widetilde{Z}_{i,j}] \right) \right]$$

$$= \frac{1}{\det(Q)} \cdot \left[- \left(\sum_{\ell=1}^{k_z} \mathbb{E}[x_i\, \widetilde{Z}_{i,\ell}] \cdot \mathbb{E}[\widetilde{Z}_{i,j}] \right) \cdot \left(\sum_{j=1}^{k_z} \mathbb{E}[\widetilde{Z}_{i,j}]^2 \right) + \left(\sum_{j=1}^{k_z} \mathbb{E}[\widetilde{Z}_{i,j}]^2 \right) \cdot \left(\sum_{\ell=1}^{k_z} \mathbb{E}[x_i \widetilde{Z}_{i,j}] \cdot \mathbb{E}[\widetilde{Z}_{i,j}] \right) \right]$$

$$= 0.$$

Hence we conclude that

$$\left.\frac{\partial \beta_{*,2}(\beta_0)}{\partial \beta_{0,2}}\right|_{\beta_0=\beta_{\mathcal{O}}} = 1.$$

.

A.3.2 Details for Example 2 (Generalized bivariate Probit)

Standard calculations (see Section 15.7.3 of Wooldridge, 2010) give:

$$\mathbb{E}_{P_{\beta_0}}[Y_i \mid X_i, Z_i] = \mathbb{E}\left[F_{U|V}(X_i'\beta_0 \mid V_i) \mid X_i, Z_i\right]$$
$$= \frac{X_i}{F_V(m(Z_i))} \cdot \int_{-m(Z_i)}^{\infty} F_{U|V}(X_i'\beta_0 \mid V_i) \cdot f_V(v)dv + \frac{1-X_i}{1-F_V(m(Z_i))} \cdot \int_{-\infty}^{-m(Z_i)} F_{U|V}(X_i'\beta_0 \mid V_i) \cdot f_V(v)dv,$$

which then gives

$$\frac{\partial s(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})}{\partial \beta_{0,2}} = \frac{\phi(\beta_2^{\mathcal{O}})}{\Phi(\beta_2^{\mathcal{O}}) \cdot (1-\Phi(\beta_2^{\mathcal{O}}))} \cdot \mathbb{E}\left[Z_i \cdot \int_{-m(Z_i)}^{\infty} f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \cdot f_V(v)dv\right].$$

Using Theorem 5 we obtain

$$\left.\frac{\partial \beta_{*,2}(\beta_0)}{\partial \beta_{0,2}}\right|_{\beta_0=\beta^{\mathcal{O}}} = \frac{1}{\phi(\beta_2^{\mathcal{O}})} \cdot \mathbb{E}[Z_i X_i]^{-1}_{(2,\bullet)} \cdot \mathbb{E}\left[Z_i \cdot \int_{-m(Z_i)}^{\infty} f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \cdot f_V(v)dv\right]$$
$$= \frac{1}{\phi(\beta_2^{\mathcal{O}})} \cdot \frac{\operatorname{Cov}\left(z_i, \int_{-m(Z_i)}^{\infty} f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \cdot f_V(v)dv\right)}{\operatorname{Cov}\left(z_i, F_V(m(Z_i))\right)}.$$

The functions $\int_{-Z_i\delta}^{\infty} f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \cdot f(v)dv$ and $F_V(m(Z_i))$ are both monotonic increasing (decreasing) in z_i when $m(Z_i)$ is monotonic increasing (decreasing). As a result, the two covariances in the above display have concordant signs and we conclude that the auxiliary IV estimator is sign consistent.

A.4 Technical Lemmas

Lemma 2. *Under the Assumptions of Theorem 5, there exists a convex and compact set $\mathcal{B}_*$ containing $\beta^{\mathcal{O}}$ such that the optimization problem*

$$\operatorname*{argmin}_{\beta \in \mathcal{B}_*} \|\gamma(\beta, \beta_0)\|_{\Omega(\beta,\beta_0)}$$

is convex for all $\beta_0 \in \mathcal{B}_*$.

A.5 Proofs

A.5.1 Proof of Lemma 1

Concavity of the log-likelihood, we have that $\widehat{\gamma}(\beta)$ is uniquely characterized by the FOC

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma,\, \alpha\right)}{\partial \gamma} Z_i = 0$$

By a mean-value expansion of the LHS in γ around $\gamma = 0$, the last display equation becomes

$$\widehat{\gamma}(\beta) = -H_n(\beta, \gamma_*(\beta))^{-1} \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \frac{d\ell\left(Y_i \,\middle|\, X_i'\beta\right)}{d\omega} Z_i \right].$$

Plugging the above into the objective function $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}$ gives the desired equivalence.

A.5.2 Proof of Theorem 3

We begin by defining the population (large n limit) analog of (10) as

$$\begin{aligned}
(\gamma(\beta), \alpha(\beta)) &= \underset{\gamma \in \mathcal{C},\, \alpha \in \mathcal{A}}{\operatorname{argmax}}\ \mathcal{L}(\beta, \gamma, \alpha), \\
\beta^* &= \left\{ \beta : \beta \in \underset{\beta \in \mathcal{B}}{\operatorname{argmin}}\ \|\gamma(\beta)\|_{\Omega_{n,\beta}} \right\}, \\
\alpha^* &= \{\alpha(\beta) : \beta \in \beta^*\}, \\
\alpha^\dagger(\beta) &= \underset{\alpha \in \mathcal{A}}{\operatorname{argmax}}\ \mathcal{L}(\beta, 0, \alpha), \\
\alpha_*^\dagger &= \left\{\alpha^\dagger(\beta) : \beta \in \beta^*\right\}.
\end{aligned}$$

The proof consists of two parts. In Part I we show that $\beta^* = \beta_0$ and $\alpha^* = \alpha_0$. In Part II we use the identification result of Part I to show consistency of $(\widehat{\beta}_{\text{AIV}}, \widehat{\alpha}_{\text{AIV}})$.

Part I: Strict concavity of the expected log-likelihood in η (Assumption 6(ii)) guarantee that $(\gamma(\beta), \alpha(\beta))$ are uniquely defined by the FOC

$$\frac{\partial \mathcal{L}(\beta, \eta)}{\partial \eta} = 0,$$

for which we have $\gamma(\beta_0) = 0$, $\alpha(\beta_0) = \alpha_0$ by Assumption 5. Suppose there exists $\check{\beta} \in \beta^*$ with $\check{\beta} \neq \beta_0$, so that

$$\begin{aligned}
\frac{\partial \mathcal{L}(\beta_0, 0, \alpha(\beta_0))}{\partial \eta} &= 0, \\
\frac{\partial \mathcal{L}(\check{\beta}, 0, \alpha(\check{\beta}))}{\partial \eta} &= 0.
\end{aligned}$$

By a mean value expansion of $\frac{\partial \mathcal{L}(\beta,0,\alpha)}{\partial \eta}$ in (β, α):

$$0 = \frac{\partial \mathcal{L}(\check{\beta}, 0, \alpha)}{\partial \eta} - \frac{\partial \mathcal{L}(\beta_0, 0, \alpha(\check{\beta}))}{\partial \eta} = \underbrace{\begin{pmatrix} \frac{\partial^2 \mathcal{L}(\widetilde{\beta},0,\widetilde{\alpha})}{\partial \gamma \partial \beta'} & \frac{\partial^2 \mathcal{L}(\widetilde{\beta},0,\widetilde{\alpha})}{\partial \gamma \partial \alpha'} \\ \frac{\partial^2 \mathcal{L}(\widetilde{\beta},0,\widetilde{\alpha})}{\partial \alpha \partial \beta'} & \frac{\partial^2 \mathcal{L}(\widetilde{\beta},0,\widetilde{\alpha})}{\partial \alpha \partial \alpha'} \end{pmatrix}}_{=A(\widetilde{\beta},0,\widetilde{\alpha})} \begin{pmatrix} \check{\beta} - \beta_0 \\ \alpha(\check{\beta}) - \alpha_0 \end{pmatrix}$$

where $(\widetilde{\beta}, \widetilde{\alpha})$ is an intermediate value between $(\check{\beta}, \alpha(\check{\beta}))$ and (β_0, α_0). The matrix $A(\widetilde{\beta}, 0, \widetilde{\alpha})$ has full rank by Assumption 6 (iii), and therefore we conclude from the previous display that

$$\check{\beta} = \beta_0, \qquad \alpha(\check{\beta}) = \alpha_0.$$

By similar arguments we have $\alpha^\dagger(\beta_0) = \alpha_0$ and thus $\alpha_*^\dagger = \alpha_0$.

Part II: show $(\widehat{\beta}_{\text{AIV}}, \widehat{\alpha}_{\text{AIV}}) = (\beta_0, \alpha_0) + o_P(1)$**.**

First define

$$\widehat{\eta}(\beta) = \underset{\eta \in \mathcal{E}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta + Z_i'\gamma,\, \alpha\right),$$
$$\eta(\beta) = \underset{\eta \in \mathcal{E}}{\operatorname{argmax}}\, \mathbb{E}\, \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta + Z_i'\gamma,\, \alpha\right),$$

Then by Pollard's convexity lemma we know that

$$\sup_{\beta \in \mathcal{B}} \sup_{\gamma \in \mathcal{E}} \left| \frac{1}{n} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta + Z_i'\gamma,\, \alpha\right) - \mathbb{E}\, \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta + Z_i'\gamma,\, \alpha\right) \right| = o_P(1).$$

Having this, we satisfy all the assumptions of Lemma B.1 in Chernozhukov and Hansen (2006), and therefore conclude

$$\sup_{\beta \in \mathcal{B}} \|\widehat{\eta}(\beta) - \eta(\beta)\| = o_P(1).$$

It directly follows that the objective function $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}$ converges uniformly to $\|\gamma(\beta)\|_\Omega$, which together with the continuity of $\|\gamma(\beta)\|_\Omega$ and β_0 being its unique minimizer over the compact set $\mathcal{B}$ (Part I) ensures that standard conditions for consistency of extremum estimators are satisfied (see, e.g., Theorem 2.1 in Newey and McFadden, 1994). We thus conclude

$$\widehat{\beta}_{\text{AIV}} = \beta_0 + o_P(1).$$

By analogous arguments we have $\sup_{(\beta,\gamma)\in(\mathcal{B},\mathcal{C})} \|\widehat{\alpha}^\dagger(\beta,\gamma) - \alpha^\dagger(\beta,\gamma)\| = o_p(1)$. Furthermore, consistency of $\widehat{\beta}_{\text{AIV}}$ and continuity of $\alpha^\dagger(\beta,\gamma)$ imply that $\alpha^\dagger(\widehat{\beta}_{\text{AIV}}, 0) = \alpha(\beta_0) + o_P(1)$. This, together with the uniform consistency of $\widehat{\alpha}^\dagger(\beta,\gamma)$, guarantees that

$$\widehat{\alpha}_{\text{AIV}} = \alpha_0 + o_P(1).$$

A.5.3 Proof of Theorem 4

The proof is in three parts. First, we show that

$$\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}} = \|s_*(\beta)\|_{W_{n,\beta}}, \tag{17}$$

with

$$\begin{aligned} s_*(\beta) &= \mathcal{L}_{n,\gamma}(\beta, 0, \alpha_0) - \mathcal{L}_{n,\gamma\alpha}(\beta, \eta_*(\beta))\, \mathcal{L}_{n,\alpha\alpha}(\beta, \eta_*(\beta))^{-1}\, \mathcal{L}_{n,\alpha}(\beta, 0, \alpha_0). \\ W_{n,\beta} &= \widetilde{H}_n(\beta, \eta_*(\beta))^{-1}\, \Omega_{n,\beta}\, \widetilde{H}_n(\beta, \eta_*(\beta))^{-1}, \end{aligned}$$

where $\eta_*(\beta) = (\gamma_*(\beta), \alpha_*(\beta))$ lies on the line between $(\widehat{\gamma}(\beta), \widehat{\alpha}(\beta))$ and $(0, \alpha_0)$. In Part II, we use the result from Part I to derive the asymptotic representation for $\widehat{\beta}_{\text{AIV}}$. In Part III, we use the result from Part II to derive the asymptotic representation for $\widehat{\alpha}_{\text{AIV}}$.

Part I: Strict concavity of the sample log-likelihood in η guarantees that $\widehat{\eta}(\beta) = (\widehat{\gamma}(\beta), \widehat{\alpha}(\beta))$ are uniquely defined by the FOC

$$\mathcal{L}_{n,\eta}(\beta, \widehat{\eta}(\beta)) = 0.$$

A mean-value expansion the above around $(\gamma, \alpha) = (0, \alpha_0)$ gives

$$\mathcal{L}_{n,\eta}(\beta, 0, \alpha_0) + \mathcal{L}_{n,\eta\eta}(\beta, \eta_*(\beta)) \cdot \widehat{\eta}(\beta) = 0 \Longrightarrow \widehat{\eta}(\beta) = -\mathcal{L}_{n,\eta\eta}(\beta, \eta_*(\beta))^{-1} \cdot \mathcal{L}_{n,\eta}(\beta, 0, \alpha_0).$$

Using the partitioned inverse formula we obtain

$$\widehat{\gamma}(\beta) = -\widetilde{H}_n(\beta, \eta_*(\beta))^{-1} \cdot s_*(\beta)$$

Plugging the above into $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,(\beta,\alpha)}}$ gives (17).

Part II: Define

$$\widehat{\beta}^\dagger := \beta_0 - \left(\widetilde{G}'\, \widetilde{W}\, \widetilde{G}\right)^{-1}\, \widetilde{G}'\, \widetilde{W}\, s(\beta), \quad s(\beta) = \mathcal{L}_{n,\gamma}(\beta, 0, \alpha_0) - \mathcal{L}_{\gamma\alpha}\, \mathcal{L}_{\alpha\alpha}^{-1}\, \mathcal{L}_{n,\alpha}(\beta, 0, \alpha_0).$$

By definition, $\widehat{\beta} := \widehat{\beta}_{\text{AIV}}$ minimizes $s_*(\beta)'\, W_{n,\beta}\, s_*(\beta)$. Therefore,

$$s_*(\widehat{\beta})'\, W_{n,\widehat{\beta}}\, s_*(\widehat{\beta}) \leq s_*(\widehat{\beta}^\dagger)'\, W_{n,\widehat{\beta}^\dagger}\, s_*(\widehat{\beta}^\dagger). \tag{18}$$

Uniform convergence of $\widehat{\eta}(\beta)$ to $\eta(\beta)$ (see proof of Theorem 3), along with consistency of $\widehat{\beta}$, implies $\eta_*(\widehat{\beta}) = (0, \alpha_0) + o_P(1)$. This, together with uniform consistency of the second

derivatives of $\mathcal{L}_n$ and $\Omega_{n,\beta}$ implies that $W_{n,\widehat{\beta}} = \widetilde{W} + o_p(1)$. Uniform convergence of $\widetilde{G}_n$ to $\widetilde{G}$ justifies the expansions

$$
\begin{aligned}
s_*(\widehat{\beta}) &= s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0) + o_P\left(\|\widehat{\beta} - \beta_0\|\right), \\
s_*(\widehat{\beta}^\dagger) &= s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0) + o_P\left(\|\widehat{\beta}^\dagger - \beta_0\|\right) = s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0) + o_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}
$$

where we have used that $\sqrt{n}(\widehat{\beta}^\dagger - \beta_0) = O_p(1)$. Plugging the expansions into (18) and using $\widetilde{W}_{n,\widehat{\beta}} = \widetilde{W} + o_P(1)$ gives for the LHS

$$
\left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0) + o_P\left(\|\widehat{\beta} - \beta_0\|\right)\right]' \widetilde{W} \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0) + o_P\left(\|\widehat{\beta} - \beta_0\|\right)\right] + R(\widehat{\beta}),
$$

with

$$
\begin{aligned}
R(\widehat{\beta}) = o_P(1) \cdot \Bigg[& s(\beta_0)'s(\beta_0) + (\widehat{\beta} - \beta_0)'\widetilde{G}'\widetilde{G}(\widehat{\beta} - \beta_0) + o_P(\|\widehat{\beta} - \beta_0\|^2) \\
& + 2\, s(\beta_0)'\widetilde{G}(\widehat{\beta} - \beta_0) + 2\, s(\beta_0)'\, o_P\left(\|\widehat{\beta} - \beta_0\|\right) + 2\,(\widehat{\beta} - \beta_0)\,\widetilde{G}'\widetilde{G}\, o_P\left(\|\widehat{\beta} - \beta_0\|\right) \Bigg] \\
= o_P & \left(\|\widehat{\beta} - \beta_0\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta} - \beta_0\| + \frac{1}{n}\right),
\end{aligned}
$$

where we have used $s(\beta_0) = O_P(1/\sqrt{n})$. Similarly, for the RHS we have

$$
\left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0) + o_P\left(\|\widehat{\beta}^\dagger - \beta_0\|\right)\right]' \widetilde{W} \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0) + o_P\left(\|\widehat{\beta}^\dagger - \beta_0\|\right)\right] + R(\widehat{\beta}^\dagger),
$$

with

$$
\begin{aligned}
R(\widehat{\beta}^\dagger) &= o_P\left(\|\widehat{\beta}^\dagger - \beta_0\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta}^\dagger - \beta_0\| + \frac{1}{n}\right) \\
&= o_P\left(\frac{1}{n}\right).
\end{aligned}
$$

Combining the previous results with the inequality (18) gives

$$
\begin{aligned}
&\left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0)\right]' \widetilde{W} \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0)\right] \\
&\quad \leq \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0)\right]' \widetilde{W} \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0)\right] + o_P\left(\|\widehat{\beta} - \beta_0\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta} - \beta_0\| + \frac{1}{n}\right).
\end{aligned} \tag{19}
$$

We now decompose $s(\beta) = A_1 + A_2$, where

$$
A_1 = \widetilde{G}(\widetilde{G}'\widetilde{W}\widetilde{G})^{-1}\widetilde{G}'\widetilde{W}s(\beta), \qquad A_2 = \left[\mathbb{I} - \widetilde{G}(\widetilde{G}'\widetilde{W}\widetilde{G})^{-1}\widetilde{G}'\widetilde{W}\right] s(\beta).
$$

Because $\widetilde{G}'\widetilde{W}A_2 = 0$, we find that the contributions of A_2 on both sides of the inequality (19) are identical and thus drop out. Also plugging in the definition of $\widehat{\beta}^\dagger$, this inequality becomes

$$\left[(\widehat{\beta}-\beta^0) + \underbrace{(\widetilde{G}'\widetilde{W}\widetilde{G})^{-1}\widetilde{G}'\widetilde{W}s(\beta_0)}_{:=L}\right]' \widetilde{G}'\widetilde{W}\widetilde{G} \times \left[(\widehat{\beta}-\beta^0) + (\widetilde{G}'\widetilde{W}\widetilde{G})^{-1}\widetilde{G}'\widetilde{W}s(\beta_0)\right] \leq o_P\left(\|\widehat{\beta}-\beta_0\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta}-\beta_0\| + \frac{1}{n}\right).$$

Because $\widetilde{G}'\widetilde{W}\widetilde{G}$ has full rank (since $\widetilde{W} > 0$ and $\text{rank}(\widetilde{G}) = k_x$) we have that

$$\begin{aligned}
\|(\widehat{\beta}-\beta_0)+L\|^2 &\leq o_P(1)\cdot\left(\|\widehat{\beta}-\beta_0\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta}-\beta_0\| + \frac{1}{n}\right) \\
&\leq o_P(1)\cdot\left(\|\widehat{\beta}-\beta_0+L\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta}-\beta_0+L\| + \|L\|^2 + \frac{1}{\sqrt{n}}\|L\| + \frac{1}{n}\right) \\
&\leq o_P\left(\frac{1}{\sqrt{n}}\right)\cdot\|\widehat{\beta}-\beta_0+L\| + o_P\left(\frac{1}{n}\right),
\end{aligned}$$

where we have used $L = O_P(1/\sqrt{n})$. Denoting $\xi_n := o_P\left(\frac{1}{\sqrt{n}}\right)\cdot\|\widehat{\beta}-\beta_0+L\|$ we can re-write the above as

$$\left(\|\widehat{\beta}-\beta_0+L\| - \xi_n\right)^2 \leq \xi_n^2 + o_P\left(\frac{1}{n}\right),$$

from which we conclude

$$\sqrt{n}(\widehat{\beta}-\beta_0) = \sqrt{n}L + o_P(1).$$

Part III: Consider the decomposition

$$\widehat{\alpha}^\dagger(\widehat{\beta}) - \alpha_0 = [\widehat{\alpha}^\dagger(\widehat{\beta}) - \widehat{\alpha}^\dagger(\beta_0)] + [\widehat{\alpha}^\dagger(\beta_0) - \alpha(\beta_0)]. \tag{20}$$

For the first term we consider the mean-value expansion:

$$\widehat{\alpha}^\dagger(\widehat{\beta}) - \widehat{\alpha}^\dagger(\beta_0) = \frac{d\widehat{\alpha}^\dagger(\beta)}{d\beta}\Big|_{\beta=\widetilde{\beta}}\cdot(\widehat{\beta}-\beta_0),$$

for $\widetilde{\beta}$ between β_0 and $\widehat{\beta}$. By the implicit function theorem, we have that in a neighbourhood of $(\beta_0, \widehat{\alpha}^\dagger(\beta_0))$

$$\frac{\partial\widehat{\alpha}^\dagger(\beta)}{\partial\beta'} = -\mathcal{L}_{n,\alpha\alpha}(\beta, 0, \widehat{\alpha}^\dagger(\beta))^{-1}\cdot\mathcal{L}_{n,\alpha\beta}(\beta, 0, \widehat{\alpha}^\dagger(\beta)).$$

Using $\widetilde{\beta} = \beta_0 + o_p(1)$ and the usual uniform convergence arguments we obtain

$$\widehat{\alpha}(\widehat{\beta}) - \widehat{\alpha}(\beta_0) = -\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta}\,(\widehat{\beta} - \beta_0) + o_p\left(\|\widehat{\beta} - \beta_0\|\right). \tag{21}$$

Furthermore, classical likelihood results give

$$\widehat{\alpha}^{\dagger}(\beta_0) - \alpha_0 = -\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{n,\alpha} + o_p(1/\sqrt{n}). \tag{22}$$

Plugging (21) and (22) into (20) gives the asymptotic representation for $\widehat{\alpha}$.

A.5.4 Proof of Theorem 5

The proof is made of three parts. In Part I we derive the formula for $\frac{d\beta_*(\beta^{\mathcal{O}})}{d\beta_0}$. In Part II we argue that $\frac{d\beta_*(\beta_0)}{d\beta_0}$ is bounded and continuous around $\beta^{\mathcal{O}}$. In Part III we finally show local sign consistency of the AIV estimator defined as the minimizer of the objective function over a suitably small closed ball around $\beta^{\mathcal{O}}$.

Part I: The function $s(\beta, \gamma, \beta_0)$ is thrice continuously differentiable, and thus by the IFT the function $\gamma(\beta, \beta_0)$ is thrice-continuously differentiable in an open set containing $(\beta, \beta_0) = (\beta^{\mathcal{O}}, \beta^{\mathcal{O}})$ with first-derivatives equal to

$$\begin{aligned}
\frac{\partial \gamma(\beta, \beta_0)}{\partial \beta'} &= [H(\beta, \gamma(\beta, \beta_0), \beta_0)]^{-1}\, G(\beta, \gamma(\beta, \beta_0), \beta_0), \\
\frac{\partial \gamma(\beta, \beta_0)}{\partial \beta_0'} &= -\,[H(\beta, \gamma(\beta, \beta_0), \beta_0)]^{-1} \cdot \frac{s(\beta, \gamma(\beta, \beta_0), \beta_0)}{\partial \beta_0'}.
\end{aligned}$$

Thrice-differentiability of $\gamma(\beta, \beta_0)$ together with Technical Lemma 2 implies that the limit of the AIV estimator $\beta^*(\beta_0)$ is characterized around $\beta^{\mathcal{O}}$ by the FOC of the minimisation in (15):

$$\Pi(\beta, \beta_0) := 2\frac{\partial \gamma(\beta, \beta_0)}{\partial \beta'}' \Omega(\beta, \beta_0)\gamma(\beta, \beta_0) + \sum_{i,j} \gamma_i(\beta, \beta_0) \cdot \gamma_j(\beta, \beta_0) \cdot \frac{\partial \Omega_{i,j}(\beta, \beta_0)}{\partial \beta} = 0, \tag{23}$$

when the estimator maximizes the objective function over the closed ball $B_{\infty,\epsilon}(\beta^{\mathcal{O}})$. We now apply the IFT to (23), where thrice-differentiability of $\gamma(\beta, \beta_0)$ implies that $\beta^*(\beta_0)$ is

twice-differentiable with

$$\begin{aligned}\frac{d\beta_*(\beta^{\mathcal{O}})}{\beta_0'} &= -\left[\frac{\partial\Pi(\beta^{\mathcal{O}},0,\beta^{\mathcal{O}})}{\partial\beta'}\right]^{-1}\frac{\partial\Pi(\beta^{\mathcal{O}},0,\beta^{\mathcal{O}})}{\partial\beta_0'} \\ &= -\left[\frac{\partial\gamma(\beta^{\mathcal{O}},\beta^{\mathcal{O}})}{\partial\beta'}'\Omega_{\mathcal{O}}\frac{\partial\gamma(\beta^{\mathcal{O}},\beta^{\mathcal{O}})}{\partial\beta'}\right]^{-1}\frac{\partial\gamma(\beta^{\mathcal{O}},\beta^{\mathcal{O}})}{\partial\beta'}'\Omega_{\mathcal{O}}\frac{\partial\gamma(\beta^{\mathcal{O}},\beta^{\mathcal{O}})}{\partial\beta_0'} \\ &= \left(G_{\mathcal{O}}'H_{\mathcal{O}}^{-1}\Omega_{\mathcal{O}}H_{\mathcal{O}}^{-1}G_{\mathcal{O}}\right)^{-1}G_{\mathcal{O}}'H_{\mathcal{O}}^{-1}\Omega_{\mathcal{O}}H_{\mathcal{O}}^{-1}\frac{\partial s(\beta^{\mathcal{O}},0,\beta^{\mathcal{O}})}{\partial\beta_0'},\end{aligned}$$

where we have used that $\gamma(\beta^{\mathcal{O}},\beta^{\mathcal{O}})=0$.

Part II: Applying the IFT twice to (23) shows, after some simple but tedious algebra, that $\frac{d^2\beta_*(\beta_0)}{(d\beta_0)^2}$ is bounded and continuous in a neighborhood of $\beta^{\mathcal{O}}$ when H,G and Ω are bounded and have singular values bounded away from zero, uniformly in (β,γ,β_0), which we assume.

Part III: We consider the Taylor expansion of $\beta_{*,k}(\beta_0)$ with respect to $\beta_{0,k}$ around $\beta^{\mathcal{O}}$:

$$\beta_{*,k}(\beta_0)=\frac{\partial\beta_{*,k}(\beta^{\mathcal{O}})}{\partial\beta_{0,k}}\cdot\beta_{0,k}+\frac{\partial^2\beta_{*,k}(\tilde{\beta})}{(\partial\beta_{0,k})^2}\cdot(\beta_{0,k})^2$$

where $\tilde{\beta}$ is an intermediate point between $(\beta_{-k}^{\mathcal{O}},\beta_{0,k})$ and $(\beta_{-k}^{\mathcal{O}},0)$, and we have used that $\beta_{0,k}^{\mathcal{O}}=0$. Multiplying the above by $\beta^{0,k}$ we obtain

$$\beta_{*,k}(\beta_0)\cdot\beta_{0,k}=\frac{\partial\beta_{*,k}(\beta^{\mathcal{O}})}{\partial\beta_{0,k}}\cdot\beta_{0,k}^2+\frac{\partial^2\beta_{*,k}(\tilde{\beta})}{(\partial\beta_{0,k})^2}\cdot\beta_{0,k}^3. \tag{24}$$

Continuity of $\frac{\partial^2\beta_{*,k}(\beta_0)}{(\partial\beta_{0,k})^2}$ implies that for an arbitrary $\varepsilon>0$ there exists a $\delta_\varepsilon>0$ such that for any $|\beta_{0,k}|<\delta_\varepsilon$ one has $\left|\frac{\partial^2\beta_{*,k}(\beta_0)}{(\partial\beta_{0,k})^2}\right|<C_\varepsilon:=\left|\frac{\partial^2\beta_{*,k}(\beta^{\mathcal{O}})}{(\partial\beta_{0,k})^2}\right|+\varepsilon$. Fixing such ε, and using that $\frac{\partial\beta_{*,k}(\beta^{\mathcal{O}})}{\partial\beta_{0,k}}>0$, we have that $\beta_{*,k}(\beta_0)\cdot\beta_{0,k}>0$ for any $\beta_{0,k}$ small enough to satisfy the requirements of the IFT in Part I and II and

$$0<|\beta_{0,k}|<\min\left\{\delta_\epsilon,\delta_\varepsilon,\frac{\partial\beta_{*,k}(\beta^{\mathcal{O}})}{\partial\beta_{0,k}}/C_\varepsilon\right\}.$$

A.5.5 Proof of Lemma 2

We want to find a convex set $\mathcal{B}_*$ containing $\beta^{\mathcal{O}}$ for which the objective function $\|\gamma(\beta,\beta_0)\|_{\Omega(\beta,\beta_0)}$ is convex in $\beta\in\mathcal{B}_*$ for all $\beta_0\in\mathcal{B}_*$. By the continuous twice-differentiability of $\gamma(\beta,\beta_0)$ and

Ω_{β,β_0} wrt (β, β_0), for every $\epsilon > 0$ there exists a δ_ϵ such that for $\|(\beta, \beta_0) - (\beta^{\mathcal{O}}, \beta^{\mathcal{O}})\|_\infty \leq \delta_\epsilon$ we have

$$\left\| \frac{\partial \|\gamma(\beta, \beta_0)\|_{\Omega_{(\beta,\beta_0)}}}{\partial\beta\, \partial\beta'} - \frac{\partial\gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})}{\partial\beta}' \Omega_{(\beta^{\mathcal{O}},\beta^{\mathcal{O}})} \frac{\partial\gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})}{\partial\beta} \right\| \leq \epsilon.$$

Denote $\mathcal{C}_\lambda$ the minimum eigenvalue of $\frac{\partial\gamma(\beta^{\mathcal{O}},\beta^{\mathcal{O}})}{\partial\beta}' \Omega_{(\beta^{\mathcal{O}},\beta^{\mathcal{O}})} \frac{\partial\gamma(\beta^{\mathcal{O}},\beta^{\mathcal{O}})}{\partial\beta}$, which is bounded away from 0 by Assumption 3. By Weyl's Inequality we have

$$\left| \lambda_{\min} \left(\frac{\partial \|\gamma(\beta, \beta_0)\|_{\Omega_{(\beta,\beta_0)}}}{\partial\beta\, \partial\beta'} \right) - \mathcal{C}_\lambda \right| \leq \epsilon$$

Choosing $\epsilon < \mathcal{C}_\lambda$ ensures that objective function is convex with respect to β over the convex set $B_{\infty,\epsilon}(\beta^{\mathcal{O}})$ for all $\beta_0 \in B_{\infty,\epsilon}(\beta^{\mathcal{O}})$, where $B_{\infty,\epsilon}(\beta^{\mathcal{O}})$ denotes a closed ball around $\beta^{\mathcal{O}}$ with respect to the ℓ_∞-norm.