



UNIVERSITÀ DI PARMA

DOTTORATO DI RICERCA IN

**"Economia e Management dell'Innovazione e della Sostenibilità"
(EMIS)**

CICLO XXXVIII°

IN CO-TUTELA con

UNIVERSIDAD DE CANTABRIA

Essays on Nonparametric Regression with Endogenous Covariates

Coordinatore:

Char.mo Prof. Beatrice Luceri

Tutori:

Chiar.mo Prof. Antonio Musolesi

Chiar.mo Prof. Alexandra Soberón Vélez

Dottorando: Davide Golinelli

Anni Accademici 2022/23 - 2024/25

Non è vero che il ricercatore insegue la verità, è la verità che insegue il ricercatore.
(Robert Musil)

Alla mia famiglia

Acknowledgments

I would like to thank from the bottom of my heart my supervisors, Antonio Musolesi and Alexandra Soberón-Vélez, for their guidance and support throughout these years. Antonio's patience, mentorship, and constant support have been a reference point for my growth, both as a researcher, as a thinker, and as a person. Alexandra's patience and help have been immense — her steady presence at every stage of this work and her careful attention to detail have left a clear mark on this thesis. Their dedication, passion, and rigor as researchers have set an example I will carry with me well beyond this doctorate.

I would like to express my sincere gratitude to the external scholars who, at different stages of this work, offered comments and suggestions that genuinely improved it: Prof. Juan Manuel Rodríguez-Poo and Prof. Jeffrey Racine.

A special mention goes to Prof. Massimiliano Mazzanti, whose support throughout this journey has been fundamental.

I would also like to thank Georgios Gioldasis and Santiago Pereda-Fernández for the stimulating discussions and the good moments shared along the way, which have left a mark that goes well beyond the academic.

To all the PhD students from my cycle, and the many researchers and professors I had the privilege of meeting at the University of Parma, the University of Ferrara, during my stay at the University of Cantabria, and at the other universities visited across these years: thank you. Sharing ideas, advice, and experiences — through the difficult moments and the good ones alike — has been one of the most rewarding parts of this journey.

To my long-standing friends, thank you for being there, year after year, with the kind of support that needs no explanation — I will always be grateful for this. To the new friends — Italian, Spanish, and from many other places — that this journey has brought into my life: I did not expect you, and you gave me more than you know — especially in the moments that counted most.

Finally, to my parents, Franco and Licia, and my sister, Chiara: everything I am able to build rests on what you have given me. Thank you, simply, for everything.

Contents

Acknowledgments	v
Riepilogo	1
Resumen	7
Summary	13
1 Nonparametric Models with Endogeneity: An Overview of Previous Works	19
1.1 Introduction	20
1.2 Dealing with Endogeneity: Full- and Limited- Information Systems	23
1.3 Functional Forms and Estimation Strategies	32
1.3.1 Endogeneity in Parametric Regression Models	33
1.3.2 Endogeneity in Nonparametric Regression Models	35
1.4 Nonparametric Techniques to Address Endogeneity	37
1.4.1 Nonparametric Instrumental Variable Regression with Regularization	38
1.4.2 Nonparametric Instrumental Variable Regression with Control Function	40
1.5 Conclusions	50
2 Continuous Endogenous Variables in Partially Linear Regression Models	53
2.1 Introduction	54
2.2 Econometric Model and Estimation Procedure	55
2.2.1 Estimation Method	57
2.2.2 Estimation Procedure	60
2.2.3 Bandwidth Selection Procedure	68
2.3 Monte Carlo Simulations	70

2.4	Conclusions	73
3	Dummy Endogenous Variables in Partially Linear Regression Models	75
3.1	Introduction	76
3.2	Econometric Model and Estimation Procedure	78
3.2.1	Estimation Procedure	82
3.2.2	Bandwidth Selection Procedure	85
3.3	Monte Carlo Simulations	86
3.3.1	Finite In-Sample Performance	87
3.3.2	Finite Out-of-Sample Performance	93
3.4	Revisiting the Knowledge Capital Production Function	96
3.5	Conclusions	105
4	Mixed-type Endogenous Variables in Partially Linear Regression Model	107
4.1	Introduction	108
4.2	Econometric Model and Estimation Procedure	112
4.2.1	Estimation Method	114
4.2.2	Estimation Procedure	116
4.2.3	Bandwidth Selection Procedure	121
4.3	Monte Carlo Simulations	122
4.4	Revisiting Knowledge Capital Production Function Estimation . .	126
4.5	Conclusions	135
5	Conclusions and Future Research	137
5.1	General Conclusions	137
5.2	Future Research	140
	Appendix	140
A	Chapter 2	141
A.1	Asymptotic Properties - Semiparametric Models with Continuous Endogenous Variables	141
B	Chapter 3	149
B.1	Estimation Procedure - Alternative First-Stage Estimators for Fit- ted Probabilities	149
B.2	Asymptotic Properties - Semiparametric Models with Discrete En- dogenous Variables	154

B.3	Empirical Illustration	159
B.3.1	Data	159
B.4	Instrumental Variables Selection	161
C	Chapter 4	163
C.1	Data	163
C.2	Instrumental Variables Selection	165
C.3	Heterogeneity across Sector-Level Productivity	167
	Bibliography	171

List of Tables

Table 1.1: Sources of endogeneity, modeling choices, and estimation approaches	31
Table 1.2: Comparison of control function estimation methods	50
Table 2.1: Finite sample performance of $\widehat{\beta}$	72
Table 3.1: Finite sample performance of $\widehat{\gamma}$ with probit $\widehat{p}$	89
Table 3.2: Finite sample performance of $\widehat{\gamma}$ with logit $\widehat{p}$	89
Table 3.3: Finite sample performance of $\widehat{\gamma}$ with single index $\widehat{p}$	90
Table 3.4: Finite sample performance of $\widehat{\gamma}$ with local likelihood logit $\widehat{p}$	90
Table 3.5: Monte Carlo results	95
Table 3.6: Parametric structural model estimates	100
Table 3.7: Nonparametric structural model estimates	102
Table 3.8: PLM-NP-IV estimates: sensitivity to Sector Bandwidth	102
Table 3.9: Estimates of input elasticity: mean and quantiles	103
Table 3.10: Elasticities of scale: mean and quantiles	103
Table 4.1: Finite sample performance of $\widehat{\beta}$: single index first stage	124
Table 4.2: Finite sample performance of $\widehat{\beta}$: local logit first stage	124
Table 4.3: Parametric structural model Estimates	127
Table 4.4: Nonparametric structural model estimates	128
Table 4.5: PLM-NP-IV estimates: sensitivity to sector bandwidth	131
Table 4.6: Estimates of input's elasticity: mean and quantiles	131
Table 4.7: Elasticities of scale: mean and quantiles	132
Table B.1: Descriptive statistics	161
Table B.2: Instrumental variables sets used in estimation	162
Table C.1: Descriptive statistics	165
Table C.2: Instrumental variables sets used in estimation	166
Table C.3: Sector-level productivity differentials: PLM-NP-IV estimates	168

List of Figures

Figure 2.1: Finite sample performance of $\widehat{m}(\cdot)$	72
Figure 3.1: Finite sample performance of $\widehat{m}(\cdot)$: probit first stage	91
Figure 3.2: Finite sample performance of $\widehat{m}(\cdot)$: logit first stage	91
Figure 3.3: Finite sample performance of $\widehat{m}(\cdot)$: single index first stage	92
Figure 3.4: Finite sample performance of $\widehat{m}(\cdot)$: local likelihood logit first stage	92
Figure 3.5: Empirical distribution of the capital elasticity: PLM-IV-NP estimates	104
Figure 3.6: Empirical distribution of the labor elasticity: PLM-IV-NP estimates	104
Figure 4.1: Finite sample performance of $\widehat{m}(\cdot)$: single index first stage	125
Figure 4.2: Finite sample properties of $\widehat{m}(\cdot)$: local logit first stage	125
Figure 4.3: Empirical distribution of the capital elasticity: PLM-NP-IV (v) estimates	133
Figure 4.4: Empirical distribution of the labor elasticity: PLM-NP-IV (v) estimates	134

Riepilogo

La crescente disponibilità di dataset microeconomici su larga scala, unita alla domanda sempre maggiore di robustezza inferenziale negli studi econometrici, ha orientato la letteratura specializzata verso specificazioni progressivamente più flessibili. In questo contesto, i modelli parzialmente lineari, che combinano una componente parametrica con una non parametrica, rappresentano un compromesso metodologicamente attraente tra flessibilità strutturale e rigore nella specificazione per gli studi empirici. La loro virtù principale risiede nell'eludere la maledizione della dimensionalità propria delle specificazioni puramente non parametriche, mantenendo al contempo la robustezza contro l'errata specificazione funzionale della componente non parametrica. Gli stimatori standard associati a questa classe di modelli sono, tuttavia, inconsistenti quando uno o più regressori presentano endogeneità. Quest'ultima, che può originarsi da variabili omesse, errori di misura nelle variabili esplicative o simultaneità tra relazioni causali, introduce una distorsione sistematica che compromette sia la validità puntuale delle stime sia l'affidabilità dell'inferenza successiva. Le strategie di correzione sviluppate per i modelli parametrici, sia tramite variabili strumentali sia attraverso l'approccio della funzione di controllo, non ammettono una trasposizione diretta al contesto semi- e non parametrico, poiché la presenza di una componente non parametrica genera complicazioni aggiuntive di natura identificativa e computazionale che richiedono un trattamento specifico. Il problema è considerevolmente più acuto quando i regressori endogeni sono discreti. In tali casi, le procedure che sfruttano la differenziabilità della funzione strutturale rispetto al

regressore endogeno, condizione tipicamente richiesta dai metodi della funzione di controllo, perdono il loro fondamento analitico. A ciò si aggiunge il fatto che la cosiddetta *forbidden regression* rende parimenti inammissibile la sostituzione diretta dei valori stimati da un primo stadio non lineare, privando il ricercatore degli strumenti di correzione più standard e richiedendo strategie di identificazione alternative. In questo quadro, l'obiettivo della presente dissertazione è duplice. Da un lato, sviluppare nuovi metodi di stima per modelli parzialmente lineari in presenza di variabili endogene, seguendo una progressione che inizia con il caso di regressori continui, affronta il caso di regressori discreti e culmina nel caso misto. Dall'altro lato, illustrare la rilevanza pratica di questi strumenti attraverso applicazioni a dati reali sull'innovazione e la produttività delle imprese italiane. A tal fine, la struttura della dissertazione è organizzata in quattro capitoli. **Il Capitolo 1** stabilisce il quadro concettuale attraverso una rassegna critica della letteratura sull'endogeneità nei modelli parametrici e non parametrici. L'analisi prende come punto di partenza la distinzione tra le tre principali fonti di endogeneità (ossia variabili omesse, errori di misura e simultaneità) e la loro diversa incidenza sulla strategia di stima più appropriata in ciascun caso. Esamina poi la dicotomia tra metodi a informazione limitata, che stimano equazione per equazione sfruttando soltanto le restrizioni identificative rilevanti, e metodi a informazione completa, che mobilitano l'intera struttura del sistema al costo di una maggiore sensibilità all'errata specificazione. Nel contesto non parametrico sorge una complicazione fondamentale: il recupero di una funzione strutturale ignota a partire dalle aspettative condizionate osservabili assume la forma di un problema inverso mal posto, in cui la soluzione non dipende in modo continuo dai dati. Il capitolo esamina dapprima la regressione non parametrica con variabili strumentali e le strategie di regolarizzazione necessarie per ripristinare la buona posizione del problema, per poi passare ai metodi della fun-

zione di controllo, esaminando gli approcci più rilevanti sviluppati per modelli non parametrici e semiparametrici parzialmente lineari. Il **Capitolo 2** analizza una procedura di stima semiparametrica per modelli parzialmente lineari in cui la variabile endogena nella componente non parametrica è di natura continua. Il modello di riferimento assume la forma

$$Y_i = m(X_{1i}, W_{1i}) + X_{2i}^\top \beta + \epsilon_i,$$

dove X_{1i} e X_{2i} sono variabili esogene, ϵ_i denota il termine idiosincratico e W_{1i} è una variabile endogena nel senso che $E(\epsilon_i | W_{1i}) \neq 0$. La procedura di stima proposta è strutturata in tre stadi sequenziali. Nel primo stadio, i residui della forma ridotta vengono recuperati tramite regressione lineare locale con una struttura kernel che gestisce congiuntamente strumenti continui e discreti attraverso la ponderazione a kernel misto proposta in J. Racine & Li (2004b). Nel secondo stadio, la trasformazione proposta in Robinson (1988) elimina la componente non parametrica, producendo uno stimatore dei minimi quadrati ordinari (OLS) per β che è $\sqrt{N}$ -consistente e asintoticamente normale sotto condizioni standard. Nel terzo stadio, la funzione ignota $m(\cdot)$ è stimata tramite integrazione marginale seguendo l'approccio proposto in Tjøstheim & Auestad (1994) e Linton & Nielsen (1995). Lo stimatore non parametrico $\hat{m}(\cdot)$ converge al tasso kernel standard ed è asintoticamente normale sotto condizioni standard. La selezione della banda segue una regola empirica gerarchica in cui il primo stadio converge più lentamente per limitare la trasmissione della varianza alle regressioni successive. Le proprietà asintotiche degli stimatori proposti sono analizzate in dettaglio nell'Appendice B. Le simulazioni Monte Carlo sotto diversi livelli di endogeneità mostrano che la distorsione, la deviazione standard e l'RMSE dello stimatore proposto per β diminuiscono sistematicamente al crescere della dimensione campionaria, anche in presenza di elevati livelli di endogeneità, confermando la stabil-

ità e la convergenza in campioni finiti dello stimatore, con risultati analoghi per la componente non parametrica $\hat{m}(\cdot)$. Il **Capitolo 3** affronta il caso in cui la variabile endogena nella componente parametrica è una variabile dummy. Il modello strutturale assume la forma

$$Y_i = m(X_{1i}) + X_{2i}^\top \beta + W_i^d \gamma + \epsilon_i,$$

dove X_{1i} e X_{2i} sono variabili esogene, ϵ_i denota il termine idiosincratico e $W_i^d \in \{0, 1\}$ è generata da una variabile latente $W_i^{d*} = g(Z_i) + u_i$, per cui l'esito binario W_i^d è determinato dal superamento di una soglia da parte di W_i^{d*} , ed è endogena in conseguenza della correlazione tra i termini di errore ϵ_i e u_i . La funzione di controllo classica è inapplicabile per tre ragioni concorrenti: i) l'errore u_i non è direttamente osservabile dalla variabile discreta W_i^d , e la sua approssimazione tramite l'inversa Mills ratio è valida soltanto sotto ipotesi distributive stringenti quali la normalità congiunta dei termini di errore ϵ_i e u_i ; ii) qualsiasi correzione CF per W_i^d rimane sensibile all'errata specificazione della forma funzionale del primo stadio, poiché una struttura dell'indice o una distribuzione dell'errore specificata in modo scorretto non eliminano completamente l'endogeneità; iii) l'introduzione diretta dei valori stimati da un primo stadio non lineare come regressore costituisce una *forbidden regression* che produce stime inconsistenti. La strategia proposta aggira questi ostacoli utilizzando la probabilità condizionata $p_i = P(W_i^d = 1 | Z_i)$ come strumento, stimata in uno stadio preliminare attraverso procedure alternative: probit, logit, stimatore a indice singolo di Klein & Spady (1993) e logit locale non parametrico di Frölich (2006). Le simulazioni Monte Carlo mostrano che gli stimatori parametrici del primo stadio presentano elevata dispersione e instabilità nei campioni di piccole dimensioni, mentre gli stimatori a indice singolo e locali producono stime ben comportate con distorsione trascurabile e varianza regolarmente decrescente. I test fuori campione su dati

indipendenti confermano che le specificazioni del modello parzialmente lineare con primo stadio non parametrico dominano uniformemente le loro controparti parametriche in termini di errore quadratico medio di previsione. L'applicazione empirica stima una funzione di produzione del capitale di conoscenza utilizzando dati sulle imprese italiane tratti dalla decima indagine Unicredit–Capitalia (2 598 imprese dopo la selezione degli strumenti), trattando l'innovazione di processo come variabile endogena. A differenza della stima OLS, che produce un effetto dell'innovazione piccolo e statisticamente non significativo ($\approx 0,020$), tutte le specificazioni a variabili strumentali restituiscono effetti positivi e significativi nell'intervallo 0,26–0,42, con errori standard più contenuti sotto il primo stadio non parametrico e un'elasticità di scala che scende da 0,93 a circa 0,81 per effetto della correzione per l'endogeneità. **Il Capitolo 4** unifica le due procedure sviluppate nei capitoli precedenti, considerando un modello parzialmente lineare in cui una variabile endogena continua W_i^c entra nella componente non parametrica accanto ai regressori esogeni X_{1i} , mentre la componente parametrica incorpora regressori binari di diversa natura: X_{2i}^d , che è esogeno, e $W_i^d \in \{0, 1\}$, che è endogeno. Il modello in esame assume la forma

$$Y_i = m(X_{1i}, W_i^c) + X_{2i}^{d\top} \beta_1 + W_i^{d\top} \beta_2 + \epsilon_i, \quad (1)$$

dove ϵ_i è il termine di errore, W_i^c è endogena nel senso che $E(\epsilon_i | W_i^c) \neq 0$ e $W_i^d \in \{0, 1\}$ è endogena in conseguenza della correlazione tra ϵ_i e il disturbo della sua equazione di selezione. La strategia di stima proposta affronta entrambe le fonti di endogeneità in modo differenziato: l'endogeneità continua è trattata tramite l'approccio della funzione di controllo, arricchendo il modello con i residui non parametrici della forma ridotta per W_i^c , mentre l'endogeneità discreta è corretta attraverso la stima a variabili strumentali con probabilità generate seguendo la procedura del Capitolo 3. La procedura a tre stadi risultante

preserva la $\sqrt{N}$ -consistenza per lo stimatore parametrico e il tasso di convergenza non parametrico ottimale per lo stimatore della funzione ignota $m(\cdot)$. Le simulazioni Monte Carlo sotto diversi livelli di endogeneità confermano il comportamento adeguato in campioni finiti degli stimatori proposti. L'applicazione empirica stima una funzione di produzione del capitale di conoscenza utilizzando 1 890 imprese manifatturiere, con strumenti per entrambe le variabili endogene. La correzione per la sola endogeneità di IP innalza sostanzialmente il suo effetto stimato (da $\approx 0,010$ a $\approx 0,34\text{--}0,40$) e riduce l'elasticità di ℓ (da $\approx 0,75$ a $\approx 0,62$). Estendendo la correzione a ℓ , questa elasticità si contrae ulteriormente a $\approx 0,53$, con l'effetto di IP che si stabilizza a $\approx 0,28\text{--}0,29$ e l'elasticità di scala che scende a $\approx 0,78$, indicativa di rendimenti moderatamente decrescenti una volta che l'endogeneità congiunta è pienamente considerata.

Resumen

La creciente disponibilidad de conjuntos de datos microeconómicos a gran escala, combinada con la demanda cada vez mayor de robustez inferencial en los estudios econométricos, ha orientado la literatura especializada hacia especificaciones progresivamente más flexibles. En este contexto, los modelos parcialmente lineales, que combinan una componente paramétrica con una no paramétrica, representan un compromiso metodológicamente atractivo entre flexibilidad estructural y rigor en la especificación para los estudios empíricos. Su virtud principal reside en eludir la maldición de la dimensionalidad propia de las especificaciones puramente no paramétricas, manteniendo al mismo tiempo la robustez frente a la especificación funcional incorrecta de la componente no paramétrica. Los estimadores estándar asociados a esta clase de modelos son, sin embargo, inconsistentes cuando uno o más regresores presentan endogeneidad. Esta última, que puede originarse en variables omitidas, errores de medición en las variables explicativas o simultaneidad entre relaciones causales, introduce un sesgo sistemático que compromete tanto la validez puntual de las estimaciones como la fiabilidad de la inferencia posterior. Las estrategias de corrección desarrolladas para los modelos paramétricos, ya sea a través de variables instrumentales o del enfoque de la función de control, no admiten una transposición directa al contexto semi- y no paramétrico, puesto que la presencia de una componente no paramétrica genera complicaciones adicionales de naturaleza identificativa y computacional que requieren un tratamiento específico. El problema es considerablemente más agudo cuando los regresores endógenos son discretos. En

tales casos, los procedimientos que explotan la diferenciabilidad de la función estructural respecto al regresor endógeno, condición típicamente requerida por los métodos de función de control, pierden su fundamento analítico. A ello se añade que la denominada *forbidden regression* vuelve igualmente inadmisibles la sustitución directa de los valores ajustados de una primera etapa no lineal, privando al investigador de las herramientas de corrección más estándar y exigiendo estrategias de identificación alternativas. En este marco, el objetivo de la presente disertación es doble. Por un lado, desarrollar nuevos métodos de estimación para modelos parcialmente lineales en presencia de variables endógenas, siguiendo una progresión que comienza con el caso de regresores continuos, aborda el caso de regresores discretos y culmina en el caso mixto. Por otro lado, ilustrar la relevancia práctica de estos instrumentos a través de aplicaciones a datos reales sobre innovación y productividad de las empresas italianas. Con este fin, la estructura de la disertación se organiza en torno a cuatro capítulos. **El Capítulo 1** establece el marco conceptual a través de una revisión crítica de la literatura sobre endogeneidad en modelos paramétricos y no paramétricos. El análisis toma como punto de partida la distinción entre las tres fuentes principales de endogeneidad (a saber, variables omitidas, errores de medición y simultaneidad) y su diferente incidencia sobre la estrategia de estimación más apropiada en cada caso. Examina luego la dicotomía entre métodos de información limitada, que estiman ecuación por ecuación aprovechando únicamente las restricciones identificativas relevantes, y métodos de información completa, que movilizan la estructura completa del sistema al coste de una mayor sensibilidad a la especificación incorrecta. En el contexto no paramétrico surge una complicación fundamental: la recuperación de una función estructural desconocida a partir de las esperanzas condicionales observables adopta la forma de un problema inverso mal planteado, en el que la solución no depende de manera continua de

los datos. El capítulo examina primero la regresión no paramétrica con variables instrumentales y las estrategias de regularización necesarias para restaurar el buen planteamiento del problema, para luego pasar a los métodos de función de control, examinando los enfoques más relevantes desarrollados para modelos no paramétricos y semiparamétricos parcialmente lineales. **El Capítulo 2** analiza un procedimiento de estimación semiparamétrico para modelos parcialmente lineales en los que la variable endógena en la componente no paramétrica es de naturaleza continua. El modelo de referencia adopta la forma

$$Y_i = m(X_{1i}, W_{1i}) + X_{2i}^\top \beta + \epsilon_i,$$

donde X_{1i} y X_{2i} son variables exógenas, ϵ_i denota el término idiosincrásico y W_{1i} es una variable endógena en el sentido de que $E(\epsilon_i | W_{1i}) \neq 0$. El procedimiento de estimación propuesto se estructura en tres etapas secuenciales. En la primera etapa, los residuos de la forma reducida se recuperan mediante regresión lineal local con una estructura kernel que maneja conjuntamente instrumentos continuos y discretos a través de la ponderación de kernel mixto propuesta en J. Racine & Li (2004b). En la segunda etapa, la transformación propuesta en Robinson (1988) elimina la componente no paramétrica, produciendo un estimador de mínimos cuadrados ordinarios (MCO) para β que es $\sqrt{N}$ -consistente y asintóticamente normal bajo condiciones estándar. En la tercera etapa, la función desconocida $m(\cdot)$ se estima mediante integración marginal siguiendo el enfoque propuesto en Tjøstheim & Auestad (1994) y Linton & Nielsen (1995). El estimador no paramétrico $\hat{m}(\cdot)$ converge a la tasa kernel estándar y es asintóticamente normal bajo condiciones estándar. La selección del ancho de banda sigue una regla empírica jerárquica en la que la primera etapa converge más lentamente para limitar la transmisión de la varianza a las regresiones posteriores. Las propiedades asintóticas de los estimadores propuestos se analizan en detalle en el Apéndice B. Las

simulaciones de Monte Carlo bajo distintos niveles de endogeneidad muestran que el sesgo, la desviación estándar y el RMSE del estimador propuesto para β disminuyen sistemáticamente a medida que crece el tamaño muestral, incluso en presencia de niveles elevados de endogeneidad, confirmando la estabilidad y la convergencia en muestras finitas del estimador, con resultados análogos para la componente no paramétrica $\hat{m}(\cdot)$. **El Capítulo 3** aborda el caso en el que la variable endógena en la componente paramétrica es una variable dummy. El modelo estructural adopta la forma

$$Y_i = m(X_{1i}) + X_{2i}^\top \beta + W_i^d \gamma + \epsilon_i,$$

donde X_{1i} y X_{2i} son variables exógenas, ϵ_i denota el término idiosincrásico y $W_i^d \in \{0, 1\}$ es generada por una variable latente $W_i^{d*} = g(Z_i) + u_i$, de modo que el resultado binario W_i^d se determina por la superación de un umbral por parte de W_i^{d*} , y es endógena como consecuencia de la correlación entre los términos de error ϵ_i y u_i . La función de control clásica es inaplicable por tres razones concurrentes: i) el error u_i no es directamente observable a partir de la variable discreta W_i^d , y su aproximación mediante la razón inversa de Mills es válida únicamente bajo supuestos distributivos estrictos como la normalidad conjunta de los términos de error ϵ_i y u_i ; ii) cualquier corrección CF para W_i^d permanece sensible a la especificación incorrecta de la forma funcional de la primera etapa, ya que una estructura del índice o una distribución del error incorrectamente especificadas no eliminan completamente la endogeneidad; iii) la introducción directa de los valores ajustados de una primera etapa no lineal como regresor constituye una *forbidden regression* que produce estimaciones inconsistentes. La estrategia propuesta sortea estos obstáculos utilizando la probabilidad condicional $p_i = P(W_i^d = 1 \mid Z_i)$ como instrumento, estimada en una etapa preliminar mediante procedimientos alternativos: probit, logit, estimador de índice

único de Klein & Spady (1993) y logit local no paramétrico de Frölich (2006). Las simulaciones de Monte Carlo muestran que los estimadores paramétricos de la primera etapa presentan elevada dispersión e inestabilidad en muestras pequeñas, mientras que los estimadores de índice único y locales producen estimaciones bien comportadas con sesgo despreciable y varianza regularmente decreciente. Las pruebas fuera de muestra sobre datos independientes confirman que las especificaciones del modelo parcialmente lineal con primera etapa no paramétrica dominan uniformemente a sus contrapartes paramétricas en términos de error cuadrático medio de predicción. La aplicación empírica estima una función de producción de capital de conocimiento utilizando datos sobre empresas italianas provenientes de la décima encuesta Unicredit–Capitalia (2 598 empresas tras la selección de instrumentos), tratando la innovación de proceso como variable endógena. A diferencia de la estimación MCO, que arroja un efecto de la innovación pequeño y estadísticamente no significativo ($\approx 0,020$), todas las especificaciones con variables instrumentales devuelven efectos positivos y significativos en el rango 0,26–0,42, con errores estándar más ajustados bajo la primera etapa no paramétrica y una elasticidad de escala que desciende de 0,93 a aproximadamente 0,81 como resultado de la corrección por endogeneidad. **El Capítulo 4** unifica los dos procedimientos desarrollados en los capítulos precedentes, considerando un modelo parcialmente lineal en el que una variable endógena continua W_i^c entra en la componente no paramétrica junto a los regresores exógenos X_{1i} , mientras que la componente paramétrica incorpora regresores binarios de distinta naturaleza: X_{2i}^d , que es exógeno, y $W_i^d \in \{0, 1\}$, que es endógeno. El modelo bajo análisis adopta la forma

$$Y_i = m(X_{1i}, W_i^c) + X_{2i}^{d\top} \beta_1 + W_i^{d\top} \beta_2 + \epsilon_i, \quad (2)$$

donde ϵ_i es el término de error, W_i^c es endógena en el sentido de que $E(\epsilon_i | W_i^c) \neq 0$ y $W_i^d \in \{0, 1\}$ es endógena como consecuencia de la correlación entre ϵ_i y la perturbación de su ecuación de selección. La estrategia de estimación propuesta aborda ambas fuentes de endogeneidad de manera diferenciada: la endogeneidad continua se trata mediante el enfoque de la función de control, enriqueciendo el modelo con los residuos no paramétricos de la forma reducida para W_i^c , mientras que la endogeneidad discreta se corrige a través de la estimación con variables instrumentales con probabilidades generadas siguiendo el procedimiento del Capítulo 3. El procedimiento en tres etapas resultante preserva la $\sqrt{N}$ -consistencia para el estimador paramétrico y la tasa de convergencia no paramétrica óptima para el estimador de la función desconocida $m(\cdot)$. Las simulaciones de Monte Carlo bajo distintos niveles de endogeneidad confirman el comportamiento adecuado en muestras finitas de los estimadores propuestos. La aplicación empírica estima una función de producción de capital de conocimiento utilizando 1 890 empresas manufactureras, con instrumentos para ambas variables endógenas. La corrección por la sola endogeneidad de IP eleva sustancialmente su efecto estimado (de $\approx 0,010$ a $\approx 0,34$ – $0,40$) y reduce la elasticidad de ℓ (de $\approx 0,75$ a $\approx 0,62$). Al extender la corrección a ℓ , esta elasticidad se contrae aún más hasta $\approx 0,53$, con el efecto de IP estabilizándose en $\approx 0,28$ – $0,29$ y la elasticidad de escala descendiendo a $\approx 0,78$, indicativa de rendimientos moderadamente decrecientes una vez que la endogeneidad conjunta se tiene plenamente en cuenta.

Summary

The growing availability of large-scale microeconomic datasets, combined with the increasing demand for inferential robustness in econometric studies, has oriented the specialized literature toward progressively more flexible specifications. In this context, partially linear models, which combine a parametric component with a nonparametric one, represent a methodologically attractive compromise between structural flexibility and rigor in specification for empirical studies. Their principal virtue lies in circumventing the curse of dimensionality inherent to purely nonparametric specifications, while maintaining robustness against functional misspecification of the nonparametric component.

Standard estimators associated with this class of models are, however, inconsistent when one or more regressors exhibit endogeneity. The latter, which may originate in omitted variables, measurement error in the explanatory variables, or simultaneity among causal relationships, introduces a systematic bias that undermines both the pointwise validity of the estimates and the reliability of subsequent inference. The correction strategies developed for parametric models, whether through instrumental variables or the control function approach, do not admit direct transposition to the semi- and non-parametric setting, since the presence of a nonparametric component generates additional complications of an identification and computational nature that require specific treatment.

The problem is considerably more acute when the endogenous regressors are discrete. In such cases, procedures that exploit the differentiability of the structural function with respect to the endogenous regressor, a condition typically required

by control function methods, lose their analytical foundation. Compounding this, the so-called *forbidden regression* renders equally inadmissible the direct substitution of fitted values from a nonlinear first stage, depriving the researcher of the most standard correction tools and calling for alternative identification strategies.

Against this backdrop, the objective of this dissertation is twofold. On the one hand, to develop new estimation methods for partially linear models in the presence of endogenous variables, following a progression that begins with the case of continuous regressors, addresses the case of discrete regressors, and culminates in the mixed case. On the other hand, to illustrate the practical relevance of these tools through applications to real data on innovation and productivity of Italian firms. To this end, the structure of this dissertation is organized around four chapters.

Chapter 1 establishes the conceptual framework through a critical review of the literature on endogeneity in parametric and nonparametric models. The analysis takes as its starting point the distinction between the three principal sources of endogeneity (i.e., omitted variables, measurement error, and simultaneity) and their different bearing on the most appropriate estimation strategy in each case. It then examines the dichotomy between limited information methods, which estimate equation by equation exploiting only the relevant identifying restrictions, and full-information methods, which mobilise the complete structure of the system at the cost of greater sensitivity to misspecification (Maddala 1992; Wooldridge 2010). In the nonparametric context, a fundamental complication arises: the recovery of an unknown structural function from the observable conditional expectations takes the form of an ill-posed inverse problem, in which the solution does not depend continuously on the data (Newey & Powell 2003). The chapter first examines nonparametric instrumental variable regression and the

regularization strategies required to restore well-posedness, then turns to control function methods, surveying the most relevant approaches developed for nonparametric and semiparametric partially linear models.

Chapter 2 analyzes a semiparametric estimation procedure for partially linear models in which the endogenous variable in the nonparametric component is continuous in nature. The reference model takes the form

$$Y_i = m(X_{1i}, W_{1i}) + X_{2i}^\top \beta + \epsilon_i,$$

where X_{1i} and X_{2i} are exogenous variables, ϵ_i denotes the idiosyncratic term, and W_{1i} is an endogenous variable in the sense that $E(\epsilon_i | W_{1i}) \neq 0$. The proposed estimation procedure is structured in three sequential stages. In the first stage, the reduced-form residuals are recovered via local linear regression with a kernel structure that jointly handles continuous and discrete instruments through the mixed kernel weighting proposed in J. Racine & Li (2004b). In the second stage, the transformation proposed in Robinson (1988) eliminates the nonparametric component, yielding an ordinary least squares (OLS) estimator for β that is $\sqrt{N}$ -consistent and asymptotically normal under standard conditions. In the third stage, the unknown function $m(\cdot)$ is estimated via marginal integration following the approach proposed in Tjøstheim & Auestad (1994) and Linton & Nielsen (1995). The nonparametric estimator $\hat{m}(\cdot)$ converges at the standard kernel rate and is asymptotically normal under standard conditions. Bandwidth selection follows a hierarchical rule of thumb in which the first stage converges more slowly to limit variance transmission to subsequent regressions. The asymptotic properties of the proposed estimators are analyzed in detail in Appendix B. Monte Carlo simulations under different levels of endogeneity show that the bias, standard deviation, and RMSE of the proposed estimator for β decrease systematically as the sample size grows, even under high levels of endogeneity, confirm-

ing the stability and finite-sample convergence of the estimator, with analogous results for the nonparametric component $\hat{m}(\cdot)$.

Chapter 3 addresses the case in which the endogenous variable in the parametric component is a dummy variable. The structural model takes the form

$$Y_i = m(X_{1i}) + X_{2i}^\top \beta + W_i^d \gamma + \epsilon_i,$$

where X_{1i} and X_{2i} are exogenous variables, ϵ_i denotes the idiosyncratic term, and $W_i^d \in \{0, 1\}$ is generated by a latent variable $W_i^{d*} = g(Z_i) + u_i$, whereby the binary outcome W_i^d is determined by whether W_i^{d*} exceeds a threshold, and is endogenous as a consequence of the correlation between the error terms ϵ_i and u_i (Gourieroux et al. 1987). The classical control function is inapplicable for three concurrent reasons: i) the error u_i is not directly observable from the discrete variable W_i^d , and its approximation via the inverse Mills ratio (J. J. Heckman 1979) is valid only under strict distributional assumptions such as joint normality of the error terms ϵ_i and u_i ; ii) any CF correction for W_i^d remains sensitive to misspecification of the first-stage functional form, since an incorrectly specified index structure or error distribution does not fully eliminate endogeneity; iii) the direct introduction of fitted values from a nonlinear first stage as a regressor constitutes a *forbidden regression* that yields inconsistent estimates (Wooldridge 2010). The proposed strategy circumvents these obstacles by using the conditional probability $p_i = P(W_i^d = 1 \mid Z_i)$ as an instrument, estimated in a preliminary stage through alternative procedures: probit, logit, single-index estimator of Klein & Spady (1993), and nonparametric local logit of Frölich (2006). Monte Carlo simulations show that parametric first-stage estimators exhibit high dispersion and instability in small samples, while single-index and local estimators produce well-behaved estimates with negligible bias and regularly decreasing variance. Out-of-sample tests on independent data confirm that partially linear

model specifications with a nonparametric first stage uniformly dominate their parametric counterparts in terms of mean squared prediction error. The empirical application estimates a knowledge capital production function using data on Italian firms from the tenth Unicredit–Capitalia survey (2,598 firms after instrument selection), treating process innovation as the endogenous variable. In contrast to the OLS estimate, which yields a small and statistically insignificant innovation effect (≈ 0.020), all instrumental variable specifications return positive and significant effects in the range 0.26–0.42, with tighter standard errors under the nonparametric first stage and a scale elasticity that falls from 0.93 to approximately 0.81 as a result of the endogenous correction.

Chapter 4 unifies the two procedures developed in the preceding chapters, considering a partially linear model in which a continuous endogenous variable W_i^c enters the nonparametric component alongside the exogenous regressors X_{1i} , while the parametric component incorporates binary regressors of different nature: X_{2i}^d , which is exogenous, and $W_i^d \in \{0, 1\}$, which is endogenous. The model under analysis takes the form

$$Y_i = m(X_{1i}, W_i^c) + X_{2i}^{d\top} \beta_1 + W_i^{d\top} \beta_2 + \epsilon_i, \quad (3)$$

where ϵ_i is the error term, W_i^c is endogenous in the sense that $E(\epsilon_i | W_i^c) \neq 0$, and $W_i^d \in \{0, 1\}$ is endogenous as a consequence of the correlation between ϵ_i and the disturbance of its selection equation. The proposed estimation strategy addresses both sources of endogeneity in a differentiated manner: continuous endogeneity is handled via the control function approach, augmenting the model with the nonparametric reduced-form residuals for W_i^c , while discrete endogeneity is corrected through instrumental variable estimation with generated probabilities following the procedure of Chapter 3. The resulting three-stage procedure preserves $\sqrt{N}$ -consistency for the parametric estimator and the optimal nonpara-

metric convergence rate for the nonparametric estimator of the unknown function $m(\cdot)$. Monte Carlo simulations under several levels of endogeneity confirm the suitable finite-sample behavior of the proposed estimators. The empirical application estimates a knowledge capital production function using 1,890 manufacturing firms, with instruments for both endogenous variables. Correcting for the endogeneity of IP alone substantially raises its estimated effect (from ≈ 0.010 to ≈ 0.34 – 0.40) and reduces the elasticity of ℓ (from ≈ 0.75 to ≈ 0.62). Extending the correction to ℓ further contracts this elasticity to ≈ 0.53 , with the effect of IP settling at ≈ 0.28 – 0.29 and the elasticity of scale declining to ≈ 0.78 , indicative of mildly decreasing returns once joint endogeneity is fully accounted for.

Chapter 1

Nonparametric Models with Endogeneity: An Overview of Previous Works

Abstract

This chapter reviews alternative approaches to addressing endogeneity in parametric and nonparametric models, examining how the choice of estimation strategy depends on the source of endogeneity, the availability of instrumental variables, and the assumptions placed on the functional form. The discussion then turns to the identification challenges specific to nonparametric settings, in particular the ill-posed inverse problem, and presents the control function approach as a tractable alternative to nonparametric instrumental variable regression with regularization.

1.1 Introduction

Endogeneity poses one of the most fundamental challenges in empirical econometrics, arising when explanatory variables are correlated with the error term and thereby invalidate standard regression assumptions (Wooldridge 2010). This correlation manifests through several mechanisms: omitted variable bias that occurs when unobserved factors influence both dependent and independent variables; simultaneity that emerges when variables are jointly determined; and measurement error that introduces systematic distortions when mismeasured regressors are correlated with true values.

The econometric literature has developed two primary frameworks for addressing endogeneity: instrumental variables and control function approaches. Instrumental variables (IV) methods achieve identification by exploiting exogenous variation in instruments—variables correlated with endogenous regressors but uncorrelated with structural errors (J. Angrist & G. Imbens 1995). Two-stage least squares (2SLS) provides the most common estimation procedure, though other estimators such as limited information maximum likelihood (LIML) and generalized method of moments (GMM) implement the same identifying assumptions (Wooldridge 2010). Control function (CF) methods address endogeneity by including first-stage residuals as additional covariates in the structural equation, thereby conditioning out the correlation between endogenous regressors and structural errors (J. J. Heckman 1978; Wooldridge 2015). While IV and CF approaches yield numerically identical estimates in linear parametric settings (Wooldridge 2010; Wooldridge 2015), they impose distinct identifying assumptions and yield different estimators in nonparametric specifications, where IV requires regularization of an ill-posed integral equation and CF requires additional separability conditions on the first-stage residuals (Newey & Powell 2003;

R. W. Blundell & Powell 2004).

This chapter reviews IV and CF approaches, starting from their parametric formulations and extending to nonparametric specifications, where the relaxation of functional form restrictions introduces new identification challenges and estimation complexities. Parametric approaches impose restrictions on functional form, such as linearity or known index structures, reducing estimation to the recovery of a finite-dimensional parameter vector. This feature ensures computational tractability and statistical efficiency when the imposed structure is correctly specified, but it entails the risk of specification bias when it is not (Wooldridge 2010). By contrast, nonparametric approaches treat the structural relationship as an unknown function to be inferred from the data, thus avoiding misspecification arising from incorrect functional form assumptions, albeit at the cost of slower convergence rates and increased sensitivity to the curse of dimensionality (Pagan & Ullah 1999). However, the nonparametric extension introduces a fundamental complication: recovering an unknown structural function from observed conditional expectations constitutes an ill-posed inverse problem, where the mapping from reduced form to structural parameters lacks continuity (Newey & Powell 2003). This requires either the use of regularization techniques with IV frameworks or the imposition of stronger identification assumptions in CF approaches (R. Blundell & Powell 2003).

Semiparametric models, and particularly partially linear models, occupy an intermediate position, treating some variables parametrically and others nonparametrically within the same specification (Robinson 1988; Powell 1994). This modularity allows the researcher to combine parametric and nonparametric components within a single model, without committing to a fully parametric or fully nonparametric structure.

The literature on Nonparametric Instrumental Variable (NPIV) regression relies

on a completeness condition to identify the structural function from instrumental variation (Newey & Powell 2003; P. Hall & Horowitz 2005; Darolles et al. 2011). Because the resulting inverse problem is ill-posed, estimation requires regularization methods to control the bias–variance trade-off (Chen 2007). In contrast, nonparametric CF methods avoid the ill-posed inverse problem by imposing conditional mean independence restrictions that permit additive separability of structural and control components (Newey, Powell, & Vella 1999). This separability enables consistent estimation of the regression function using kernel methods (Su & Ullah 2008) or series approximations (Newey, Powell, & Vella 1999; Pinkse 2000), where the nonparametric component can be estimated via backfitting (Hastie & Tibshirani 1990; Kim et al. 1999; Martins-Filho & Yao 2012) or marginal integration techniques (Tjøstheim & Auestad 1994; Newey 1994; Linton & Nielsen 1995). Extensions to partially linear models have also been proposed (Martins-Filho & Yao 2012; Delgado & Parmeter 2014), addressing endogeneity in both components of the model or in the nonparametric function alone, for the case of continuous endogenous variables.

The choice between IV and CF approaches entails fundamental trade-offs that differ across parametric and nonparametric settings. In parametric models, IV methods rely on the assumption of mean independence between instruments and structural errors, while CF approaches impose the stronger requirement that the first-stage residual fully captures the endogenous correlation (Wooldridge 2010; Wooldridge 2015). Under correct specification, both approaches achieve $\sqrt{n}$ -consistency. In nonparametric models, however, this distinction becomes substantially more consequential. IV methods retain the weaker mean independence assumption but face the challenge of ill-posed inverse problems, necessitating regularization techniques and resulting in slower convergence rates (Newey & Powell 2003). By contrast, nonparametric CF approaches circumvent

ill-posedness by imposing full independence between instruments and structural errors—a considerably stronger requirement—while recovering standard nonparametric convergence rates (R. Blundell & Powell 2003; G. W. Imbens & Newey 2009). The semiparametric partially linear models analyzed in this thesis provide a balanced framework to navigate these trade-offs, combining parametric specification for the endogenous component with nonparametric flexibility in the treatment of control variables.

The structure of this chapter proceeds as follows. Section 1.2 reviews the concept of endogeneity and IV, and introduces the structural representation framework. Section 1.3.2 examines functional form assumptions, including parametric and nonparametric models in the presence of endogeneity. Section 1.4 focuses on nonparametric techniques, with particular attention to IV regression under regularization methods and CF approaches in triangular models, and also provides a brief overview of both seminal and recent contributions to the CF literature. Finally, Section 1.5 concludes.

1.2 Dealing with Endogeneity: Full- and Limited- Information Systems

Endogeneity constitutes one of the most pervasive challenges in empirical econometrics, arising whenever explanatory variables are correlated with unobserved determinants of the outcome, thereby violating the exogeneity condition required for consistent estimation. Formally, endogeneity occurs when there is correlation between the regressor and the error term that leads standard estimators, such as ordinary least squares, to yield inconsistent parameter estimates, as the estimated coefficients conflate the causal effect of interest with spurious correlation driven by unobserved confounding factors.

The sources of endogeneity are well understood and can be broadly classified into three primary categories. First, omitted variables constitute a fundamental source, arising when unobserved factors influence both the explanatory variable and the outcome. For instance, in the estimation of returns to education, unobserved ability affects both schooling decisions and earnings, leading OLS to overstate the causal effect of education (Card 1999). Similarly, in macroeconomic growth studies, institutional quality or historical factors may simultaneously shape investment and long-run growth, generating spurious correlations (Acemoglu, Johnson, et al. 2001). Second, measurement error arises when regressors are observed with error, particularly when such errors are correlated with the true variable or with unobserved determinants of the outcome. Typical examples include misreported income in labor and health studies, or mismeasured productivity and capital stock in firm-level analyses (Syverson 2011; Wooldridge 2010). Finally, simultaneity represents a distinct source of endogeneity, occurring when explanatory variables and outcomes are jointly determined, thereby giving rise to feedback effects. In macroeconomic growth models, for example, investment and GDP growth may be simultaneously determined, as higher investment promotes growth while higher expected growth attracts investment (R. W. Blundell & Powell 2004). In microeconomic settings, variables such as prices and quantities, or wages and labor supply, are jointly determined through the interaction of economic agents (J. D. Angrist & Krueger 1991).

The appropriate econometric response depends fundamentally on the underlying economic structure generating the data and on whether the endogenous relationship of interest can be meaningfully represented as a single autonomous causal link, a triangular simultaneous system, or as part of a fully simultaneous system of jointly determined variables. Following Maddala (1992), this distinction is closely related to the choice between limited-information and full-

information approaches, which pertains to the extent of the system exploited during estimation rather than to the statistical technique employed. Limited-information methods estimate structural parameters equation-by-equation, relying only on the identifying restrictions pertinent to the equation of interest and without requiring complete specification of all behavioral relationships or of the joint distribution of endogenous variables. By contrast, full-information methods estimate all structural parameters jointly, fully exploiting the system structure, cross-equation restrictions, and the joint distribution of the error terms.

The motivation for adopting limited-information methods arises when the object of interest is a specific causal relationship and the researcher seeks to remain agnostic about auxiliary components of the economic system. When endogeneity originates from omitted variables or measurement error, the empirical analysis typically centers on a single structural equation that retains an economically meaningful interpretation in isolation. In such contexts, limited-information estimation delivers consistent parameter estimates without requiring potentially restrictive or fragile assumptions regarding other relationships that are not directly pertinent to the research question.

Full-information methods are motivated in settings where the researcher is interested in multiple structural parameters across several equations, where the entire system is well-understood and plausibly correctly specified, and where the potential efficiency gains from joint estimation justify the additional modeling burden. The implications of this choice are substantial. Limited-information methods remain consistent for the parameters of the equation of interest even if other equations in the system are misspecified, provided that the identifying assumptions for that equation are satisfied; however, this robustness typically comes at the cost of efficiency. By contrast, full-information methods achieve asymptotic efficiency by exploiting the complete system structure, but any misspecification

in a single equation may propagate throughout the system, potentially rendering all parameter estimates inconsistent (Wooldridge 2010; Maddala 1992).

Within the limited-information framework, IV methods constitute the canonical approach to estimation. This methodology offers a fundamental econometric solution to endogeneity by exploiting variables that are correlated with the endogenous regressor yet orthogonal to the structural error term. This dual requirement, commonly referred to as relevance and exogeneity, enables the instrument to isolate the exogenous variation in the endogenous regressor, thereby allowing for the consistent estimation of the causal parameter of interest (Wooldridge 2010). Consider a single structural equation of the form

$$Y_i = m(W_i, X_i) + \epsilon_i, \quad (1.1)$$

where Y_i denotes the outcome variable, W_i is a continuous endogenous regressor such that $\mathbb{E}(\epsilon_i | W_i) \neq 0$, X_i represents a set of exogenous covariates, ϵ_i is the structural error term, and $m(\cdot)$ is an unknown function to be estimated. The full set of instruments is given by $Z_i = (Z_{1i}, X_i)$, where Z_{1i} denotes external exogenous variables while X_i comprises the exogenous variables in the structural equation. Identification relies on two key assumptions. First, *instrument validity* (or the exclusion restriction), requires that $\mathbb{E}(\epsilon_i | Z_i) = 0$, implying that the instruments Z_i affect Y_i only through W_i and are uncorrelated with the structural error ϵ_i . Second, *instrument relevance* requires that Z_i and W_i exhibit sufficient correlation, conditional on X_i , to ensure identification. These assumptions are inherently untestable and must be justified on economic or institutional grounds. Classic examples include the use of quarter of birth as an instrument for education (J. D. Angrist & Krueger 1991), and settler mortality as an instrument for institutional quality (Acemoglu, Johnson, et al. 2001).

Under these conditions, two-stage least squares (2SLS) provides consistent es-

estimates of the structural parameters without requiring full specification of the joint distribution of (Y_i, W_i) or of auxiliary relationships beyond those necessary to construct valid instruments. The function $m(\cdot)$ may be specified parametrically or left unrestricted, a distinction that is explored in the subsequent section on functional form specification.

CF methods provide an alternative limited-information approach that relies on an additional structural assumption. Like IV methods, CF approaches use instruments to address endogeneity, but differ in their implementation: specifically, they model the reduced-form relationship between the endogenous variable and the instruments, and incorporate the resulting first-stage residual as an additional regressor in the structural equation (Wooldridge 2010).

The control function approach is grounded in the conditional independence assumption

$$\mathbb{E}(\epsilon_i \mid Z_i, u_i) = \mathbb{E}(\epsilon_i \mid u_i) = \lambda(u_i), \quad (1.2)$$

where u_i denotes the error term in the reduced-form equation linking W_i to the instruments Z_{1i} and exogenous covariates X_i , and ϵ_i is the structural error. This condition implies that the structural error ϵ_i is mean-independent of the endogenous regressor W_i , conditional on the reduced-form error u_i and exogenous covariates Z_i . Notably, this assumption is stronger than the standard IV exclusion restriction, as it requires that all sources of endogeneity in W_i operate exclusively through its correlation with u_i , such that conditioning on u_i renders W_i effectively exogenous.

Under this framework, the model can be represented as a triangular simultaneous system of the form

$$\begin{cases} Y_i = m(W_i, X_i) + \epsilon_i, \\ W_i = g(Z_i) + u_i, \quad \text{with } \text{Corr}(\epsilon_i, u_i) \neq 0. \end{cases} \quad (1.3)$$

Without the conditional independence assumption in (1.2), a triangular system of the form (1.3) cannot be obtained, rendering consistent estimation of the structural equation substantially more challenging (Li & J. S. Racine 2007). The system is triangular in the sense that W_i is determined prior to Y_i , and Y_i does not enter the equation for W_i . When $\text{Corr}(\epsilon_i, u_i) = 0$, the system is triangular recursive and OLS yields consistent estimates of the structural equation. By contrast, when $\text{Corr}(\epsilon_i, u_i) \neq 0$, the system becomes triangular simultaneous, as unobserved factors jointly influence both W_i and Y_i , thereby generating endogeneity. This setup differs from a fully simultaneous system, in which Y_i and W_i appear in each other's equations, a case explicitly ruled out by the triangular structure. It is important to distinguish triangular simultaneous systems from full simultaneous equation models. In system (1.3), causality operates unidirectionally from W_i to Y_i , and endogeneity arises exclusively from the correlation between error terms. By contrast, fully simultaneous equation models are characterized by bidirectional feedback, whereby Y_i and W_i are jointly determined and each variable enters as a regressor in the equation of the other, generating endogeneity through structural interdependence (Wooldridge 2010). The triangular structure therefore constitutes a special case which, although it still necessitates the use of IV due to correlated disturbances, circumvents the additional identification challenges associated with reverse causality.

The distinction between triangular simultaneous and triangular recursive systems is essential for determining the presence of endogeneity and, consequently, the appropriate estimation strategy (Maddala 1983; Maddala 1992). In system (1.3), the correlation $\text{Corr}(\epsilon_i, u_i) \neq 0$ characterizes the model as triangular simultaneous, indicating that common unobserved factors jointly influence both W_i and Y_i , even in the absence of feedback from Y_i to W_i . This correlation between error terms implies that W_i is endogenous in the structural equation for Y_i , i.e.,

$\mathbb{E}(\epsilon_i | W_i, X_i) \neq 0$, rendering OLS inconsistent and motivating the use of CF methods for consistent estimation. A canonical example is the estimation of returns to education, where unobserved ability simultaneously affects both schooling decisions and labor market earnings; such ability enters both error terms, inducing correlation and thereby making education endogenous in the wage equation (Card 1999).

By contrast, a triangular system is classified as recursive when the error terms are uncorrelated, $\text{Corr}(\epsilon_i, u_i) = 0$, implying that unobserved factors affecting W_i are independent of those influencing Y_i . Under this condition, W_i is exogenous in the structural equation, that is, $\mathbb{E}(\epsilon_i | W_i, X_i) = 0$, and OLS yields consistent estimates without the need for IV. In such cases, the system can be estimated sequentially: first by estimating the reduced-form equation for W_i , and subsequently by estimating the structural equation for Y_i conditional on the observed values of W_i (Wooldridge 2010). A canonical example arises in randomized controlled trials, where treatment assignment W_i is determined by a lottery; the randomization mechanism ensures that unobserved factors affecting assignment are independent of unobserved determinants of the outcome, thereby generating a recursive structure that is amenable to consistent estimation via OLS.

Under assumption (1.2), the structural equation in the triangular simultaneous system (1.3) can be reformulated by incorporating the reduced-form error u_i as an additional regressor:

$$Y_i = m(W_i, X_i) + r(u_i) + \eta_i, \quad (1.4)$$

where $\eta_i = \epsilon_i - r(u_i)$, and $\mathbb{E}(\eta_i | W_i, X_i, u_i) = 0$ under appropriate regularity conditions (Newey, Powell, & Vella 1999; R. W. Blundell & Powell 2004). The term u_i captures the endogenous component of W_i that is correlated with the structural error ϵ_i , and the additive representation in (1.4) is equivalent to the

conditional mean restriction in 1.2. The triangular structure ensures that estimation can proceed on an equation-by-equation basis, without requiring joint solution or full-system specification. In linear settings with a linear first stage and homoskedastic errors, CF and 2SLS estimators are algebraically equivalent. The key distinction, however, lies in the explicit modeling of the endogenous regressor's dependence on the instruments: CF methods incorporate the first-stage residuals, $\widehat{u}_i$, as a proxy for u_i , thereby making the underlying conditional independence structure transparent and facilitating extensions to nonlinear contexts where 2SLS may be inconsistent. For example, in the estimation of labor supply elasticities with discrete choice frameworks, or in the analysis of treatment effects with binary outcomes, CF methods retain consistency, whereas applying 2SLS directly to nonlinear models generally does not. The CF framework accommodates flexible, potentially nonlinear specifications of both $g(\cdot)$ and $r(\cdot)$, as discussed in the subsequent section. Importantly, although the equations in (1.3) form a system, estimation remains within a limited-information framework: each equation is treated sequentially, and no joint likelihood or cross-equation covariance structure is imposed.

When simultaneity with feedback is present, such that endogenous variables are jointly determined and appear in each other's equations, fully simultaneous system-of-equations models become necessary. In this context, both limited-information methods (such as equation-by-equation 2SLS) and full-information methods (such as Full-Information Maximum Likelihood (FIML) and three-stage least squares (3SLS)) may be employed. However, full-information approaches require stronger assumptions regarding the correct specification of the entire system (Wooldridge 2010), and their detailed treatment lies beyond the scope of this work.

Table 1.1 summarizes the relationship between sources of endogeneity, struc-

tural representations, identifying assumptions, and the corresponding estimation strategies.

Table 1.1: Sources of endogeneity, modeling choices, and estimation approaches

System Type	Source of Endogeneity & Structure	Estimation Approach
Single Equation	Omitted variables or measurement error. No feedback from Y_i to W_i. $Y_i = m(W_i, X_i) + \epsilon_i$ where $\mathbb{E}(\epsilon_i W_i, X_i) \neq 0$ but $\mathbb{E}(\epsilon_i Z_i) = 0$. Assumptions: Instrument exogeneity $\mathbb{E}(\epsilon_i Z_i) = 0$ and instrument relevance (sufficient correlation between Z_i and W_i conditional on X_i).	Limited-information methods: 2SLS. Instruments used directly via moment condition $\mathbb{E}[(Y_i - m(W_i, X_i)) \cdot Z_i] = 0$ without specifying reduced form or error structure.
Triangular Simultaneous System	Omitted variables or measurement error with explicit error correlation but no feedback from Y_i to W_i. $\begin{cases} Y_i = m(W_i, X_i) + \epsilon_i, \\ W_i = g(Z_i) + u_i \end{cases}$ with $\text{Corr}(\epsilon_i, u_i) \neq 0$. Assumptions: Instrument exogeneity $\mathbb{E}(u_i Z_i, X_i) = 0$ and conditional independence $\mathbb{E}(\epsilon_i Z_i, u_i) = \mathbb{E}(\epsilon_i u_i) = \lambda(u_i)$	Limited-information methods: Control function (CF) or 2SLS. Sequential estimation: first estimate reduced form for W_i, then include $\hat{u}_i$ in structural equation. CF exploits recursive structure and extends naturally to nonlinear models where 2SLS may be inconsistent. In linear models with homoskedastic errors, CF and 2SLS are algebraically equivalent.
Triangular Recursive System	No error correlation and no feedback from Y_i to W_i. $\begin{cases} Y_i = m(W_i, X_i) + \epsilon_i, \\ W_i = g(Z_i) + u_i \end{cases}$ with $\text{Corr}(\epsilon_i, u_i) = 0$. Implication: $\mathbb{E}(\epsilon_i W_i, X_i) = 0$ (no endogeneity).	Standard methods: OLS is consistent. Sequential estimation without instruments: estimate W_i equation first, then estimate Y_i equation by OLS.
Fully Simultaneous System	Structural simultaneity with mutual determination and feedback between endogenous variables. $\begin{cases} Y_{1i} = m_1(Y_{2i}, X_{1i}) + \epsilon_{1i}, \\ Y_{2i} = m_2(Y_{1i}, X_{2i}) + \epsilon_{2i} \end{cases}$ Assumptions: Exclusion restrictions (each equation excludes exogenous variables appearing in the other) and order/rank conditions for identification.	Limited-information: 2SLS equation-by-equation using excluded exogenous variables as instruments. Robust to misspecification in other equations. Full-information: 3SLS or FIML. Achieves asymptotic efficiency by exploiting cross-equation restrictions and the joint error distribution, but misspecification in any equation propagates to all parameter estimates.

1.3 Functional Forms and Estimation Strategies

Once the structural representation has been specified, estimation requires additional assumptions regarding functional form. These assumptions determine how the underlying economic relationships are approximated in the empirical model and interact closely with the chosen econometric framework.

Parametric approaches impose a known functional form on the structural relationship. While the linear specification constitutes the canonical case, parametric models can accommodate nonlinearities through explicit functional form choices (such as polynomial expansions, interaction terms, or index structures) while preserving the identification conditions underpinning standard IV methods. Identification is typically achieved via 2SLS or likelihood-based procedures, relying on instruments that satisfy both relevance and exogeneity. In this context, relevance corresponds to a rank condition on $\mathbb{E}[Z_i'W_i]$, whereas exogeneity requires that $\mathbb{E}[Z_i'\epsilon_i] = 0$.

Nonparametric IV models relax functional form restrictions by treating the structural relationship as an unknown function to be estimated. While this added flexibility mitigates the risk of functional form misspecification, it renders identification substantially more demanding. In addition to requiring valid instruments, nonparametric IV approaches rely on sufficient variation in W_i conditional on Z_i , as well as on a completeness condition that ensures the injectivity of the associated conditional expectation operator. These requirements make nonparametric IV estimation considerably more data-intensive and highly sensitive to the curse of dimensionality, thereby underscoring a fundamental trade-off between flexibility, interpretability, and empirical feasibility.

1.3.1 Endogeneity in Parametric Regression Models

A common encountered specification arises in parametric models, where the unknown functions $m(\cdot)$ and $g(\cdot)$ are assumed to be linear, and the error terms enter additively.

Considering the following parametric model

$$Y_i = \beta W_i + \epsilon_i \quad (1.5)$$

where Y_i denotes the dependent variable, W_i represents continuous endogenous regressors, and ϵ_i is the error term. In this setting, endogeneity arises when W_i is correlated with ϵ_i , i.e., $\text{Cov}(W_i, \epsilon_i) \neq 0$. This correlation leads to biased and inconsistent parameter estimates, thereby necessitating appropriate econometric corrections (Wooldridge 2010).

A range of methods has been developed to address endogeneity, among which IV estimation remains one of the most widely used. Within this framework, consistent estimation is achieved by exploiting additional information provided by instruments that induce exogenous variation in the endogenous regressors. This typically entails specifying a system of equations in which the instruments serve to isolate the component of variation in the endogenous variables that is orthogonal to the structural error term. The credibility of this approach hinges critically on the validity of the instruments and the correct specification of the underlying model.

Alternatively, the CF approach addresses endogeneity by explicitly modeling through residual-based corrections. This method entails estimating the endogenous regressors in a first-stage regression, recovering the associated residuals, and incorporating these residuals as additional regressors in the structural equation to account for endogeneity (see Wooldridge (2010), Wooldridge (2015) among others).

1.3.1.1 Instrumental Variables Approach

The IV method is most commonly implemented via a two-stage least-squares (2SLS) approach. In the first stage, we regress W_i on Z_i using OLS, yielding the fitted values

$$\widehat{W}_i = \widehat{\pi}Z_i. \quad (1.6)$$

In the second stage, these fitted values are substituted into the structural equation, replacing the endogenous regressor

$$Y_i = \beta \widehat{W}_i + \epsilon_i. \quad (1.7)$$

The resulting 2SLS estimator can be expressed as

$$\widehat{\beta}_{2SLS} = \left(\widehat{W}^\top \widehat{W} \right)^{-1} \widehat{W}^\top Y, \quad (1.8)$$

where $\widehat{W}$ and Y denote the vectors of fitted values and outcomes stacked over observations $i = 1, \dots, n$.

1.3.1.2 Control Functions Approach

In the formulation emphasized by Wooldridge (2015), the reduced-form equation for the endogenous regressor is given by

$$W_i = \pi Z_i + u_i, \quad (1.9)$$

where the first-stage residual is estimated as $\widehat{u}_i = W_i - \widehat{\pi}Z_i$. By construction, $\widehat{u}_i$ is mean-independent of the instruments Z_i . The structural error ϵ_i must satisfy the restriction in (1.2), which ensures that conditioning on the residual purges the endogenous regressor of its correlation with unobserved determinants.

Identification then relies on specifying a functional form for the CF, commonly expressed as

$$\mathbb{E}(\epsilon_i \mid u_i) = \delta u_i, \quad (1.10)$$

together with correct specification of the first-stage equation, so that the residual captures the component of W_i correlated with the structural disturbance. In addition, sufficient variation in the instruments is required to identify both the structural parameters and the control term.

In the second stage, the obtained residual, $\widehat{u}_i$, serves as the CF. The corresponding estimator is obtained by including $\widehat{u}_i$ as an additional regressor, yielding

$$\widehat{\eta}_{CF} = ([W \ \widehat{u}]^\top [W \ \widehat{u}])^{-1} [W \ \widehat{u}]^\top Y, \quad (1.11)$$

where $\widehat{\eta}_{CF} = (\widehat{\beta}, \widehat{\delta})^\top$, and W , $\widehat{u}$, and Y denote the vectors stacked over observations $i = 1, \dots, n$.

1.3.2 Endogeneity in Nonparametric Regression Models

Nonparametric regression models provide a flexible framework for analyzing complex relationships without imposing restrictive functional form assumptions. In this section, we consider the following nonparametric regression model

$$Y_i = m(W_i) + \epsilon_i, \quad (1.12)$$

where $m(\cdot)$ is an unknown function, W_i is a $q_W \times 1$ vector of endogenous variables, and ϵ_i is the error term.

A central challenge in estimating $m(W_i)$ arises when W_i is endogenous. Identification typically relies on the condition

$$\mathbb{E}(\epsilon_i | Z_i) = \mathbb{E}[Y_i - m(W_i) | Z_i] = 0, \quad (1.13)$$

where Z_i is a $q_Z \times 1$ vector of IV. This identification condition leads to the integral equation

$$\mathbb{E}(Y_i | Z_i) = \int m(W_i) f(W_i | Z_i) dW_i, \quad (1.14)$$

which links the structural function $m(\cdot)$ to observable quantities through a Fred-

holm integral equation of the first kind (Kress 1999). Inverting this relationship to recover $m(W_i)$ from $\mathbb{E}(Y_i|Z_i)$ is not straightforward: small perturbations in the observable conditional expectation can translate into large deviations in the recovered structural function, rendering nonparametric IV estimation inherently ill-posed (Newey & Powell 2003).

1.3.2.1 Ill-Posed Inverse Problem

In econometric analysis, such inverse problems arise when researchers seek to recover underlying structural relationships from observed data. These problems are termed *well-posed* if they satisfy Hadamard's criteria: (i) existence of a solution, (ii) uniqueness of that solution, and (iii) continuous dependence of the solution on the data (Hadamard 2014). In many practical applications, particularly in nonparametric settings, these conditions fail to hold, giving rise to *ill-posed* problem.

The ill-posedness of inverse problems in nonparametric econometrics typically stems from two main sources. First, lack of identification may imply that multiple structural functions are observationally equivalent, thereby violating uniqueness. In linear IV models, identification is associated with the invertibility of a finite-dimensional matrix. In contrast, the nonparametric analogue involves a compact operator, whose inversion is inherently problematic. This led to the second issue: instability. Small perturbations in the data, arising from measurement error or sampling variability, can generate large fluctuations in the estimated function. Such sensitivity reflects the ill-conditioned nature of the inverse problem, whereby minor discrepancies in the inputs translate into substantial distortions in the recovered structural relationship. A prominent example arises in NPIV models, where estimation requires inverting a conditional expectations operator that is typically ill-conditioned, thereby exacerbating instability (P. Hall

& Horowitz 2005; Darolles et al. 2011).

To obtain meaningful and robust estimates, it is therefore necessary to impose additional structure or constraints that mitigate the effects of ill-posedness. A range of econometric approaches has been developed to address this challenge, balancing the flexibility of nonparametric models with the need for stable inference. On the one hand, NPIV methods exploit instruments to identify structural functions but require regularization techniques to stabilize the inversion (Newey & Powell 2003; Horowitz 2011; Centorrino et al. 2017). On the other hand, CF approaches rely on auxiliary equations to correct for endogeneity, thereby offering an alternative route to identification that avoids direct inversion of ill-conditioned operators (J. J. Heckman & Robb 1985; Wooldridge 2015; Newey, Powell, & Vella 1999; Pinkse 2000; R. W. Blundell & Powell 2004; Su & Ullah 2008).

1.4 Nonparametric Techniques to Address Endogeneity

To address endogeneity in nonparametric models, a variety of methodologies have been proposed in the literature, most notably IV techniques and CF approaches. In this chapter, particular emphasis is placed on the CF framework developed by Newey, Powell, & Vella (1999) as a primary method for mitigating endogeneity. While both IV and CF approaches rely on exclusion restrictions for identification in nonlinear settings, CF methods may yield efficiency gains by exploiting additional structure in the conditional distribution.

This section reviews both NPIV and CF approaches to handling endogeneity, with a specific focus on the latter, which forms the methodological foundation for the subsequent chapters.

1.4.1 Nonparametric Instrumental Variable Regression with Regularization

Nonparametric instrumental variable (NPIV) methods aim to recover the structural function $m(\cdot)$ by solving a conditional moment restriction of the form $\mathbb{E}[Y_i | Z_i] = \mathbb{E}[m(W_i) | Z_i]$.

The validity of instruments in this setting requires stronger conditions than in the parametric case. In particular, instruments Z_i must satisfy both *exogeneity* and *completeness*. The exogeneity condition requires that the instrument be uncorrelated with the structural error term,

$$\mathbb{E}[\epsilon_i | Z_i] = 0 \quad (1.15)$$

ensuring that Z_i does not convey information about unobserved determinants of Y_i , in line with the standard IV assumption.

More critically, identification requires a *completeness condition* with respect to the endogenous regressor W_i . Completeness guarantees that the mapping from the structural function $m(\cdot)$ to the observable conditional expectation $\mathbb{E}[Y_i | Z_i]$ is injective, thereby allowing for the identification of $m(W_i)$. Formally, let $\mathcal{H}$ denote a function space (e.g., L^2). The completeness condition requires that

$$\mathbb{E}[h(W_i) | Z_i] = 0 \quad \text{a.s.} \Rightarrow h(W_i) = 0 \quad \text{a.s.}, \quad \forall h \in \mathcal{H}. \quad (1.16)$$

This condition implies that the conditional distribution of W_i given Z_i is sufficiently rich such that no non-zero function of W_i is orthogonal to all functions of Z_i . In contrast to the linear model, where identification is ensured by a rank condition (i.e., full column rank of the matrix of instruments), completeness is considerably more demanding and generally untestable in practice (Dufour 2003). Completeness may fail in finite samples or when the support of Z_i is limited, resulting in a non-invertible conditional expectation operator and lack of identifi-

cation. While exogeneity can sometimes be supported through economic reasoning or tested indirectly via overidentifying restrictions, completeness typically relies on structural assumptions or functional form restrictions, making it a central challenge in NPIV estimation.

Beyond identification, estimation is further complicated by the ill-posed nature of the inverse problem, which necessitates the use of regularization techniques. Regularization introduces constraints or penalties into the estimation procedure to stabilize the inversion and prevent overfitting (Tikhonov et al. 1943; Newey & Powell 2003). In particular, it ensures the existence and uniqueness of the solution, improves numerical stability by mitigating ill-conditioning, and supports consistency by controlling the trade-off between bias and variance as the regularization parameter $\lambda_n \rightarrow 0$ at an appropriate rate. Additionally, regularization often enforces smoothness restrictions that are essential for identification (see Centorrino et al. (2017) for further discussion in additive NPIV contexts).

While NPIV methods provide a powerful framework for addressing endogeneity, their implementation can be computationally demanding and sensitive to specification choices. CF approaches offer an alternative by recasting endogeneity as an omitted variable problem and imposing additional structure to avoid the ill-posed inverse problem. Specifically, CF methods rely on a triangular simultaneous system in which first-stage residuals control for the endogenous component in the structural equation (Newey, Powell, & Vella 1999; Su & Ullah 2008). This framework is particularly appealing due to its computational tractability and its natural compatibility with kernel-based estimation techniques, which are employed in the subsequent chapters.

1.4.2 Nonparametric Instrumental Variable Regression with Control Function

The CF approach provides an alternative strategy for addressing endogeneity by explicitly modeling the dependence between the endogenous regressor and the structural error within a nonparametric framework.

In parametric settings, the CF method typically relies on specific distributional assumptions, most commonly joint normality of the structural and first-stage errors, to justify the inclusion of first-stage residuals as additional regressors (J. J. Heckman 1978; Rivers & Vuong 1988). The parametric literature often assumes a linear relationship of the form $\mathbb{E}[\epsilon_i | Z_i, u_i] = \gamma u_i$, where u_i denotes the error term of the auxiliary regression of the endogenous variable W_i on the instruments Z_i (Wooldridge 2010). This linearity assumption, together with exclusion restrictions ensuring that Z_i affects Y_i only through W_i , enables consistent two-step estimation without requiring a fully specified functional form for the structural equation.

The nonparametric CF framework relaxes these distributional assumptions while preserving the two-step estimation structure. In the first stage, residuals $\hat{u}_i$ are obtained from a nonparametric projection of the endogenous variable W_i on instruments Z_i and exogenous covariates X_i . In the second stage, these residuals are included as additional arguments in a fully nonparametric regression of the outcome Y_i on $(X_i, W_i, \hat{u}_i)$. The residuals capture the endogenous component of W_i that is correlated with the structural error, and under appropriate conditional independence restrictions, their inclusion as control variates effectively corrects for endogeneity in the estimation of the structural function.

1.4.2.1 Foundational Nonparametric Control Function Approach

Two principal modeling frameworks are typically considered in the nonparametric CF literature, differing in whether exogenous covariates enter the structural equation alongside the endogenous regressor. The first, proposed by Newey, Powell, & Vella (1999), builds on the triangular system in (1.3) and the conditional independence assumption in (1.2) — which, as noted by Li & J. S. Racine (2007), is necessary to obtain such a structure.

Direct estimation of (1.3) using standard procedures yields inconsistent estimates of the structural function due to the endogeneity of W_i (i.e., $\mathbb{E}[\epsilon_i | W_i] \neq 0$). In particular, ignoring endogeneity invalidates the moment conditions underlying conventional nonparametric estimators.

Taking conditional expectations on both sides of the structural equation and exploiting the conditional independence assumption in (1.2), it follows that the conditional mean no longer coincides with the structural function of interest. Specifically, we obtain

$$\begin{aligned}\mathbb{E}[Y_i | W_i, Z_i, u_i] &= m(W_i, Z_{1i}) + \mathbb{E}[\epsilon_i | W_i, Z_i, u_i] \\ &= m(W_i, Z_{1i}) + r(u_i),\end{aligned}\tag{1.17}$$

where the second equality follows from 1.2. This expression highlights that the observed conditional expectation incorporates an additional component, $r(u_i)$, capturing the endogenous variation. Consequently, failure to account for this term leads to biased estimation of the structural function.

The conditional mean independence assumption in (1.2) is crucial for ensuring the additive separability in (1.17). Absent this condition, the structural error ϵ_i would remain dependent on the instruments even after conditioning on u_i , leading to a nonseparable interaction between the control function and the structural function. The assumption therefore guarantees that the endogenous component

enters through $r(u_i)$, remaining disentangled from the structural relationship $m(W_i, Z_{1i})$. This triangular decomposition is fundamental, as it permits the separate identification and estimation of the structural function and the nuisance CF.

The resulting model can be written as

$$Y_i = m_0(W_i, Z_{1i}, u_i) + \eta_i, \quad (1.18)$$

where $m_0(W_i, Z_{1i}, u_i) = m(W_i, Z_{1i}) + r(u_i)$, and $\eta_i = \epsilon_i - r(u_i)$ denotes the new error term.

The function $m_0(W_i, Z_{1i}, u_i)$ admits an additive structure in which $m(W_i, Z_{1i})$ captures the CF correction. However, this decomposition introduces a standard identification issue in additive nonparametric models: the individual components are not separately identified without normalization. In particular, for any constant c , the representation

$$m_0(W_i, Z_{1i}, u_i) = [m(W_i, Z_{1i}) + c] + [r(u_i) - c] \quad (1.19)$$

is observationally equivalent to (1.18). This location indeterminacy necessitates the imposition of normalization restrictions, such as $\mathbb{E}[r(u_i)] = 0$ or fixing the value of $r(u_i)$ at a specific point.

Estimation under the CF approach proceeds in two stages. In the first stage, since u_i is unobserved, the continuous endogenous regressor W_i is projected onto the instrument set Z_i to estimate the reduced-form function $g(Z_i)$ and obtain residuals that serve as proxies for the unobserved CF. Newey, Powell, & Vella (1999) propose implementing this step using sieve approximations based on basis functions $(\psi_1, \dots, \psi_{J_n})$, where the number of terms J_n increases with the sample

size at an appropriate rate. The resulting series estimator is given by

$$\widehat{g}(Z_i) = \sum_{j=1}^{J_n} \hat{\beta}_j \psi_j(Z_i), \quad (1.20)$$

where the coefficients $\hat{\beta}_j$ are obtained via least squares regression of W_i on the basis functions $\psi_j(Z_i)$. Common choices for $\psi_j(\cdot)$ include polynomial bases, splines, and wavelets. The first-stage residuals are then constructed as

$$\widehat{u}_i = W_i - \widehat{g}(Z_i), \quad (1.21)$$

which provide consistent estimates of the unobserved component u_i as $n \rightarrow \infty$ and $J_n \rightarrow \infty$ at suitable rates.

In the second stage, the structural function $m(W_i, Z_{1i})$ is estimated nonparametrically by regressing Y_i on $(W_i, Z_{1i}, \widehat{u}_i)$. Given the additive structure in (1.17), the structural function can be separated from the nuisance CF using methods such as marginal integration or backfitting procedures. Marginal integration recovers $m(W_i, Z_{1i})$ by averaging out the contribution of the control function:

$$\widehat{m}(W_i, Z_{1i}) = \int \widehat{m}_0(W_i, Z_{1i}, \widehat{u}_i) d\widehat{F}_{\widehat{u}}(\widehat{u}_i), \quad (1.22)$$

where $\widehat{F}_{\widehat{u}}$ denotes the empirical distribution of the first-stage residuals $\widehat{u}_i$. This integration effectively marginalizes over the distribution of the CF, leaving only the structural component up to the location normalization. Alternatively, backfitting procedures iteratively estimate each additive component through partial regression until convergence.

Newey, Powell, & Vella (1999) show that, under appropriate regularity conditions (including smoothness conditions on $m(\cdot)$, $g(\cdot)$, and $r(\cdot)$, suitable growth rates for J_n , and regularity conditions on the basis functions), the resulting series estimator is consistent. Convergence rates depend on the smoothness of the unknown functions and the dimensionality of the regressors; while the full function typi-

cally converges at slower nonparametric rates, $\sqrt{n}$ -consistent estimation may be achieved for certain functionals.

This framework represents a significant extension of earlier parametric CF methods by relaxing functional form assumptions while maintaining computational tractability and avoiding the ill-posed inverse problem inherent in NPIV estimation. However, this advantage comes at the cost of stronger identifying assumptions: in particular, the CF approach relies on the conditional mean independence condition in (1.2), whereas NPIV methods can be identified under the weaker condition $\mathbb{E}[\epsilon_i | Z_i] = 0$. This trade-off between identification strength and computational feasibility is central to the choice between CF and IV approaches in nonparametric settings.

A second structural specification commonly considered in this literature takes the form

$$\begin{cases} Y_i = m(W_i) + \epsilon_i, \\ W_i = g(Z_i) + u_i, \end{cases} \quad (1.23)$$

where, relative to system (1.3), the structural function depends solely on the endogenous regressor W_i and Z_i contains only external exogenous variables. Pinkse (2000) studies identification and estimation in this setting, proposing a two-stage estimator under the full independence assumption $Z_i \perp (\epsilon_i, u_i)$. This condition is stronger than the conditional mean independence in (1.2), and its stringency is directly tied to the absence of exogenous covariates from the nonparametric structural function: since X_i does not enter $m(\cdot)$ to provide additional identifying variation, the instruments must be clean in the stronger sense of full independence to nonparametrically separate $m(W_i)$ from $r(u_i)$ without contamination from higher-order dependence between Z_i and the error terms. Under this assumption, the first-stage residual $\widehat{u}_i = W_i - \widehat{g}(Z_i)$ serves as a valid control variable,

yielding the additive decomposition

$$\mathbb{E}[Y_i | W_i, u_i] = m(W_i) + r(u_i), \quad r(u_i) = \mathbb{E}[\epsilon_i | u_i].$$

Both unknown functions are then approximated via series expansions,

$$m(w) = \sum_{l=0}^{L_n} \beta_{wl} e_{wl}(w), \quad r(u_i) = \sum_{l=0}^{L_n} \beta_{ul} e_{ul}(u_i),$$

where $e_{wl}(\cdot)$ and $e_{ul}(\cdot)$ are basis functions such as polynomials or splines, β_{wl} and β_{ul} are expansion coefficients, and L_n grows with the sample size at an appropriate rate. Substituting into the conditional mean yields

$$m_0(w) = \beta_0 + \sum_{l=1}^{L_n} \beta_{wl} e_{wl}(w) + \sum_{l=1}^{L_n} \beta_{ul} e_{ul}(u_i), \quad (1.24)$$

where $\beta_0 = \beta_{w0} + \beta_{u0}$. Since equation (1.24) is linear in the coefficients, the second stage reduces to ordinary least squares after substituting $\widehat{u}_i$ for u_i . While computationally convenient, the approach relies on the stringent full independence assumption and, by virtue of its global approximation strategy, may inadequately capture local features of $m(\cdot)$ in regions of high curvature or heterogeneity.

In contrast to the global nature of series methods in Pinkse (2000), Su & Ullah (2008) adopt a fundamentally different approach based on local smoothing techniques. Working with the specification in (1.3), which incorporates additional exogenous covariates Z_{1i} , they employ kernel-based local polynomial estimation in both stages. Importantly, their identification strategy relaxes the strong independence assumption used by Pinkse (2000), requiring instead the weaker conditions $\mathbb{E}[u_i | Z_i] = 0$ and $\mathbb{E}[\epsilon_i | Z_i, u_i] = \mathbb{E}[\epsilon_i | u_i]$. This weaker assumption allows for a broader class of data-generating processes, permitting correlation between instruments and structural errors through the endogenous regressor while still ensuring identification.

The estimation procedure follows a two-stage kernel regression strategy. In the

first stage, W_i is regressed on Z_i using local polynomial smoothing of order p_1 , with kernel K_1 and bandwidth h_1 , yielding estimates $\widehat{g}(Z_i)$ and residuals $\widehat{u}_i = W_i - \widehat{g}(Z_i)$. These residuals serve as proxies for the unobserved CF. In the second stage, the augmented regression function $m_0(w, z_1, u) = m(w, z_1) + r(u)$ is estimated via local polynomial smoothing of order p_2 applied to Y_i on $(W_i, Z_{1i}, \widehat{u}_i)$, using kernel K_2 and bandwidth h_2 .

Under the conditional independence assumption, the structural function $m(w, z_1)$ is recovered through marginal integration to eliminate the nuisance component $r(u)$:

$$\widehat{m}(w, z_1) = \int \widehat{m}_0(w, z_1, u) dQ(u), \quad (1.25)$$

where Q is a weighting function. When $\mathbb{E}[\epsilon_i] = 0$, full identification can be achieved by setting $Q(u) = \widehat{F}_u(u)$, the empirical distribution of $\widehat{u}_i$.

Under standard regularity conditions on the kernel functions, bandwidth sequences, and the smoothness of the underlying functions, Su & Ullah (2008) establish asymptotic normality:

$$\sqrt{nh_2^{d_w+d_1}} \left(\widehat{m}(w, z_1) - m(w, z_1) - h_2^{p_2+1} B(w, z_1) \right) \xrightarrow{d} N(0, V(w, z_1)), \quad (1.26)$$

where $B(w, z_1)$ denotes the asymptotic bias, $V(w, z_1) = \kappa_0 \sigma_\epsilon^2(w, z_1) / f_{W, Z_1}(w, z_1)$ with $\kappa_0 = \int K_2^2(u) du$, and d_w, d_1 represent the dimensions of W_i and Z_{1i} , respectively.

The comparison between these approaches highlights a fundamental trade-off. Kernel-based methods offer superior local adaptivity and are better suited to capturing heterogeneous effects across the support of the data. However, they are particularly vulnerable to the curse of dimensionality: as the combined dimension $d_w + d_1 + 1$ increases, convergence rates deteriorate rapidly due to data sparsity. By contrast, series-based methods achieve more favorable convergence rates in higher-dimensional settings, typically of order $O_p(n^{-r/(2r+d)})$ where r de-

notes smoothness and d dimensionality, but at the cost of reduced flexibility due to their global approximation structure. Consequently, the choice between these approaches involves balancing local adaptivity against scalability in higher dimensions, depending on the complexity and dimensionality of the empirical setting.

Martins-Filho & Yao (2012) extend the triangular simultaneous system in (1.3) to a semiparametric partially linear framework, allowing endogenous regressors to enter both the nonparametric and linear components of the structural equation:

$$Y_i = m(W_{1i}, X_{1i}) + W_{2i}^\top \beta + \varepsilon_i,$$

with reduced-form system

$$W_i = g(Z_i) + U_i, \quad \mathbb{E}(U_i | Z_i) = 0,$$

where $W_i = (W_{1i}, W_{2i})$ is a vector of endogenous variables, X_i collects exogenous covariates, Z_i denotes the instruments, $U_i = (u_{1i}, u_{2i})$ is the vector of first-stage disturbances, and ε_i is the structural error. The CF assumption $\mathbb{E}(\varepsilon_i | Z_i, U_i) = \mathbb{E}(\varepsilon_i | U_i)$ ensures that conditioning on U_i fully accounts for the endogeneity of W_i . Under this assumption, the structural equation admits the representation

$$\mathbb{E}(Y_i | W_{1i}, X_{1i}, U_i) = m(W_{1i}, X_{1i}) + W_{2i}^\top \beta + h(U_i),$$

where $h(U_i) = \mathbb{E}(\varepsilon_i | U_i)$ is the CF.

Estimation proceeds in three steps. First, the reduced-form equation is estimated using Nadaraya–Watson kernel regression, yielding residuals $\hat{U}_i = (\hat{u}_{1i}, \hat{u}_{2i})$. Second, a Robinson-type partialling-out is applied, conditioning on $(W_{1i}, X_{1i}, \hat{U}_i)$. In this step, $\hat{u}_{1i}$ controls for the endogeneity of W_{1i} in the nonparametric component, while $\hat{u}_{2i}$ controls for the endogeneity of W_{2i} in the linear component through the transformation $W_{2i} - \mathbb{E}(W_{2i} | W_{1i}, X_{1i}, \hat{U}_i)$ which removes the endogenous variation in W_{2i} conditional on the first-stage residuals. Third, the parame-

ter vector β is obtained via OLS applied to the partialled-out regression, yielding a closed-form estimator that avoids numerical optimization. The nonparametric component $m(\cdot)$ is subsequently recovered through a backfitting procedure applied to the augmented function $g(W_{1i}, X_{1i}, \hat{U}_i) = m(W_{1i}, X_{1i}) + h(\hat{U}_i)$, which separates the structural function $m(\cdot)$ from the CF $h(\cdot)$ by integrating out the residuals over their empirical distribution. Monte Carlo evidence reported by Martins-Filho & Yao (2012) indicates that this estimator exhibits superior finite-sample performance relative to sieve-based alternatives, particularly in settings characterized by strong endogeneity.

Delgado & Parmeter (2014) extend the triangular simultaneous system in 1.23 and propose a marginal integration estimator for a semiparametric partially linear model with endogeneity in the nonparametric component. The structural equation is the following,

$$Y_i = m(W_i) + X_i^\top \beta + \epsilon_i,$$

where W_i is an endogenous regressor, X_i is a vector of exogenous covariates, ϵ_i is the structural error term, and $m(\cdot)$ is the structural function of interest.

The model assumes

$$\mathbb{E}[u_i|Z_i] = 0, \quad \mathbb{E}[\epsilon_i|u_i, Z_i] = \mathbb{E}[\epsilon_i|u_i],$$

which implies that conditional on u_i , the structural error ϵ_i is mean-independent of the instruments Z_i .

Under these conditions, the model admits a CF representation of the form

$$Y_i = m(W_i) + X_i^\top \beta + h(u_i) + v_i,$$

where $h(u_i) = \mathbb{E}[\epsilon_i|u_i]$ and v_i is a residual orthogonal to all regressors. This additive decomposition separates the structural function $m(\cdot)$ from the CF $h(\cdot)$, thereby facilitating estimation via marginal integration.

The estimation procedure is constructed in three steps. First, u_i is estimated from the reduced-form equation using local linear regression. Second, the augmented regression

$$Y_i = m_0(W_i, \hat{u}_i) + X_i^\top \beta + v_i,$$

is considered, where $m_0(W_i, \hat{u}_i) = m(W_i) + h(\hat{u}_i)$. A Robinson (1988)-type transformation is then applied by estimating the conditional expectations $\mathbb{E}[Y_i|W_i, \hat{u}_i]$ and $\mathbb{E}[X_i|W_i, \hat{u}_i]$ via local linear regression, and regressing the residualized outcome on the residualized covariates to obtain β . The function $m_0(\cdot)$ is subsequently estimated nonparametrically. In the final step, the structural function is recovered through marginal integration:

$$\hat{m}(W_i) = \frac{1}{n} \sum_{j=1}^n \hat{m}_0(W_i, \hat{u}_j).$$

which averages out $\widehat{u}_i$. All nonparametric components are estimated using kernel-based local linear methods, with bandwidths selected via least-squares cross-validation. Notably, this approach is computationally straightforward, as it avoids iterative optimization and does not require inversion of integral operators. Monte Carlo evidence indicates that the marginal integration estimator performs comparably to the backfitting estimator of Martins-Filho & Yao (2012), and may outperform it in small samples, particularly under strong endogeneity.

Table 1.2 summarizes the main methodological choices in constructing CF estimators. Kernel and series approaches involve a trade-off between flexibility and sensitivity to tuning parameters in the first stage, while backfitting and marginal integration differ in the degree of structure imposed in the second stage. Each approach reflects a distinct balance between computational complexity, smoothness assumptions, and robustness to misspecification.

Table 1.2: Comparison of control function estimation methods

Stage / Method	Key Conditions and Strengths	Limitations and Trade-offs
First stage: Estimation of the reduced-form equation and CF residuals		
Kernel Regression	Nonparametric estimation of $\mathbb{E}[W_i Z_i]$ via local polynomial smoothing. Superior local adaptivity; captures heterogeneous effects across the support of covariates. Compatible with weaker conditional mean independence ($\mathbb{E}[u_i Z_i] = 0$).	Suffers acutely from the curse of dimensionality: convergence deteriorates rapidly when the combined dimension of regressors exceeds three. Bandwidth selection critically affects bias-variance trade-off.
Series Approximation	Approximates $\mathbb{E}[W_i Z_i]$ via basis expansions (e.g., polynomials, splines, wavelets) with J_n terms growing with n . Achieves better convergence rates in higher dimensions, typically $O_p(n^{-r/(2r+d)})$. Computationally efficient through least-squares implementation.	Global approximation sacrifices local flexibility; may poorly capture regions of high curvature or heterogeneity. Choice of basis and truncation parameter J_n affects performance: underfitting if J_n too small, overfitting if too large. Pinkse's version requires the stringent full independence assumption $Z_i \perp\!\!\!\perp (\epsilon_i, u_i)$.
Second stage: Recovery of the structural function $m(\cdot)$		
Backfitting	Iteratively estimates each additive component ($m(\cdot)$ and $r(\cdot)$) by partial regression. Directly exploits the additive separability $m_0 = m + r$. Monte Carlo evidence shows strong finite-sample performance, particularly under strong endogeneity (Martins-Filho & Yao 2012).	Relies on the additive structure imposed by the CF assumption. May converge slowly depending on smoothing parameters and starting values. Estimation quality sensitive to bandwidth choices at both stages.
Marginal Integration	Recovers $m(W_i)$ by averaging $\widehat{m}_0(W_i, \hat{u}_i)$ over the empirical distribution of $\hat{u}_i$: $\widehat{m}(W_i) = n^{-1} \sum_j \widehat{m}_0(W_i, \hat{u}_j)$. Avoids iterative optimization. Computationally straightforward with no inversion of integral operators. Comparable performance to backfitting, with potential advantages in small samples (Delgado & Parmeter 2014).	Still requires the additive decomposition $m_0(W_i, u_i) = m(W_i) + h(u_i)$ for separate identification of $m(\cdot)$. Location normalization needed (e.g., $\mathbb{E}[r(u_i)] = 0$). Integration step is computationally intensive in large samples.

1.5 Conclusions

This chapter reviews the principal nonparametric approaches to addressing endogeneity in econometric models. The literature has developed two main methodological frameworks: NPIV methods with regularization and CF approaches. NPIV methods achieve identification through completeness conditions and address the resulting ill-posed inverse problem via regularization techniques (Newey & Powell 2003; P. Hall & Horowitz 2005; Darolles et al. 2011). In contrast, CF methods circumvent the ill-posed inverse problem by imposing conditional mean independence, which enables an additive decomposition of the structural and control components. Within the CF framework, first-stage estimation of $\mathbb{E}[W | Z]$

is typically carried out using either series approximations (Newey, Powell, & Vella 1999; Pinkse 2000) or kernel-based local polynomial methods (Su & Ullah 2008; Delgado & Parmeter 2014). In the second stage, the structural function is recovered through procedures such as backfitting (Martins-Filho & Yao 2012) or marginal integration (Su & Ullah 2008; Delgado & Parmeter 2014). Semiparametric partially linear models, as developed by Martins-Filho & Yao (2012) and Delgado & Parmeter (2014), provide an intermediate framework that combines parametric structure for selected components with nonparametric flexibility for the structural function. These approaches retain much of the robustness of nonparametric methods while alleviating, to a considerable extent, the identification and computational challenges inherent in the full NPIV framework.

Chapter 2

Continuous Endogenous Variables in Partially Linear Regression Models

Abstract

This chapter introduces a semiparametric estimation method designed to address continuous endogenous variables in regression models. Under standard regularity conditions, the proposed approach ensures consistent and efficient estimation of both parametric and nonparametric components. Moreover, it provides flexible and robust tools for empirical researchers confronting endogeneity within semiparametric frameworks. The asymptotic properties of the resulting estimator are formally established, and the theoretical findings are further supported by Monte Carlo simulations aimed at evaluating its finite-sample performance.

2.1 Introduction

Endogeneity constitutes a fundamental identification problem in econometric modeling, arising when explanatory variables are correlated with unobserved determinants of the outcome variable. Such correlation compromises the validity of standard estimation procedures by inducing systematic bias and inconsistency in parameter estimates. This issue becomes particularly pronounced in semiparametric settings, where the objective is to preserve flexibility in functional form while ensuring credible identification of causal effects.

Over recent decades, substantial progress has been achieved in the development of consistent estimation methodologies for semiparametric models involving continuous endogenous variables. Seminal contributions by Newey & Powell (2003) and R. W. Blundell & Powell (2004) laid the theoretical groundwork for addressing endogeneity within flexible specifications through CF approaches and IV techniques. These methods typically exploit triangular or recursive system structures, in which the endogenous regressor is modeled as a function of exogenous instruments.

This chapter focuses on partially linear models in which continuous endogenous variables enter the model nonparametrically. A prominent example arises in the estimation of production functions, where a firm's output depends on continuously measured inputs such as capital and labor. These inputs are commonly endogenous due to simultaneity in factor demand decisions or their correlation with unobserved productivity shocks (D. A. Akerberg et al. 2015; Levinsohn & Petrin 2003). The semiparametric framework accommodates flexible, potentially nonlinear representations of the production technology without imposing restrictive functional form assumptions, while the CF approach addresses endogeneity by conditioning on residuals derived from reduced-form equations.

The central insight underlying these methods is that, within triangular system structures exhibiting additive separability in unobservables, the conditional expectation of the outcome given the endogenous regressor and a suitably constructed control variable identifies the structural function of interest. This strategy preserves the flexibility inherent in semiparametric modeling while introducing sufficient structure to distangle exogenous variation in the endogenous regressor from variation driven by unobserved factors. Identification relies on the availability of valid exclusion restrictions (i.e., variables that influence the endogenous regressor but do not directly affect the outcome), as well as on functional form conditions such as monotonicity or strict exogeneity of instruments. The chapter proceeds by establishing the theoretical framework for partially linear models with continuous endogenous regressors entering the nonparametric component. We characterize the identification conditions under which the structural function can be consistently estimated, develop the associated multi-step estimation procedure, and derive its asymptotic properties (see Appendix A). The remainder of the chapter is organized as follows: Section 2.2 presents the general econometric specification under consideration and examines the consistent estimation of partially linear models with continuous endogenous variables in the nonparametric component. Section 2.3 reports the results of Monte Carlo simulations designed to evaluate the finite-sample performance of the proposed estimator. Finally, 2.4 concludes.

2.2 Econometric Model and Estimation Procedure

In this chapter, we consider the consistent estimation of the following partially linear model

$$Y_i = m(X_{1i}, W_{1i}) + X_{2i}^\top \beta + \epsilon_i, \quad (2.1)$$

where Y_i denotes an observable scalar dependent variable, X_{1i} and X_{2i} are $(q_{x_1} \times 1)$ and $(q_{x_2} \times 1)$ vectors of exogenous variables, respectively, and W_{1i} is a $(q_{w_1} \times 1)$ vector of endogenous variables. The function $m(\cdot) : \mathbb{R}^{q_{x_1} + q_{w_1}} \rightarrow \mathbb{R}$ is assumed to be an unknown smooth function, β is a finite-dimensional parameter vector, and ϵ_i denotes the disturbance term.

When W_{1i} is exogenous, such that $E(\epsilon_i|W_{1i}) = 0$, conventional estimation methods, such as that proposed in Robinson (1988), yield consistent estimates of β . However, in the presence of endogeneity, i.e., $E(\epsilon_i|W_{1i}) \neq 0$, these standard approaches generally lead to inconsistent estimators.

To address this issue, the primary objective of this chapter is to examine the consistent estimation of a semiparametric regression model in which the endogenous variable enters as a continuous argument within a smooth function.

Prior to detailing the estimation procedures, we introduce the kernel functions employed to smooth over both continuous and categorical variables. Let $Z_i = (Z_i^c, Z_i^d)$ denote a $(q_z \times 1)$ vector of IV where Z_i^c represents the first q_z^c continuous instruments and Z_i^d the remaining $q_z^d = q_z - q_z^c$ discrete instruments. For each discrete component Z_{ir}^d , we assume that it takes values in a finite set $Z_{ir}^d \in 0, 1, \dots, c_r - 1$ for $r = 1, \dots, q_z^d$. Let $z = (z^c, z^d) \in \mathbb{R}^{q_z}$ denote a generic point, and let $f(z) = f(z^c, z^d)$ denote the joint density of (Z_i^c, Z_i^d) . Furthermore, define $\mathcal{D} = \prod_{r=1}^{q_z^d} 0, 1, \dots, c_r - 1$ as the support of Z_i^d .

For the continuous components Z_i^c , we define the product kernel as

$$K_h(Z_i^c, z^c) = \prod_{s=1}^{q_z^c} \frac{1}{h_s} k\left(\frac{Z_{is}^c - z_s^c}{h_s}\right), \quad (2.2)$$

where $k(\cdot)$ is a univariate kernel function and $h = (h_1, \dots, h_{q_z^c})$ is a vector of bandwidth parameters.

For the discrete regressors, we focus on the case in which no natural ordering is

present (e.g., binary or unordered categorical variables), and define

$$\ell(Z_{ir}^d, z_r^d, \lambda) = \begin{cases} 1 & \text{if } Z_{ir}^d = z_r^d \\ \lambda & \text{if } Z_{ir}^d \neq z_r^d, \end{cases} \quad (2.3)$$

where λ is a smoothing parameter that lies within the interval $[0, 1]$. Notably, $\ell(Z_{ir}^d, z_r^d, 0)$ reduces to an indicator function, whereas $\ell(Z_{ir}^d, z_r^d, 1)$ corresponds to a uniform weighting scheme. Let $\mathbb{1}(Z_{ir}^d \neq z_r^d)$ denote the indicator function that equals one if $Z_{ir}^d \neq z_r^d$ and zero otherwise. Define $d_{Z_i, z} = \sum_{r=1}^{q_z^d} \mathbb{1}(Z_{ir}^d \neq z_r^d)$ as the number of mismatched components. The product kernel for the discrete variables can then be written as

$$L(Z_i^d, z^d, \lambda) = \prod_{r=1}^{q_z^d} \ell(Z_{ir}^d, z_r^d, \lambda) = \mathbb{1}^{1-d_{Z_i, z}} \lambda^{d_{Z_i, z}}, \quad (2.4)$$

where $\lambda = (\lambda_1, \dots, \lambda_{q_z^d})$ is the vector of smoothing parameters. Finally, let $\phi = (h, \lambda) = (h_1, \dots, h_{q_z^c}, \lambda_1, \dots, \lambda_{q_z^d})$ collect all smoothing parameters.

The joint kernel weight function for both continuous and discrete variables is defined as

$$\mathcal{W}_\phi(Z_i, z) = \prod_{s=1}^{q_z^c} \frac{1}{h} k\left(\frac{Z_{is}^c - z_s^c}{h}\right) \prod_{r=1}^{q_z^d} \ell(Z_{ir}^d, z_r^d, \lambda) = K_h(Z_i^c, z^c) L(Z_i^d, z^d, \lambda). \quad (2.5)$$

2.2.1 Estimation Method

In this section, we examine the consistent estimation of partially linear models in which the nonparametric component contains continuous endogenous regressors, while the available instruments may be either continuous or discrete. In this setting, endogeneity implies that conventional semiparametric estimators are unable to recover the structural effect of the endogenous regressor, since that regressor is correlated with unobserved disturbances entering the outcome equation. To address this problem, we adopt a *triangular* system that explicitly mod-

els the data-generating process of the endogenous variable and thereby provides a foundation for the construction of an appropriate CF.

Triangular systems are especially suitable for semiparametric estimation because they disentangle the determination of the endogenous regressor from the structural relationship of interest in a manner that captures the dependence between the endogenous regressor and the structural disturbance. The essential insight of the CF approach is that the component of the endogenous variable correlated with the structural error can be isolated through the first-stage equation (Newey, Powell, & Vella 1999). Once an appropriate function of this component into the outcome equation, the remaining variation is rendered conditionally exogenous, thereby allowing for consistent estimation of the partially linear structural relationship. This reasoning has been extended to nonparametric and semiparametric environments by Su & Ullah (2008), Martins-Filho & Yao (2012), and, by Delgado & Parmeter (2014).

More specifically, we consider the following triangular simultaneous equations model

$$\begin{cases} Y_i = m(X_{1i}, W_i^c) + X_{2i}^\top \beta + \epsilon_i, \\ W_i^c = g(Z_i) + u_i, \end{cases} \quad E(\epsilon_i | Z_i) = 0, \quad E(\epsilon_i | Z_i, u_i) = E(\epsilon_i | u_i), \quad (2.6)$$

where X_{1i} and X_{2i} are $(q_{x_1} \times 1)$ and $(q_{x_2} \times 1)$ vectors of continuous exogenous variables, W_i^c is a $(q_w \times 1)$ vector of endogenous continuous variables, and Z_i is a $(q_z \times 1)$ vector of excluded instruments that may be either continuous or discrete. In addition, $g(\cdot)$ is a vector of unknown smooth functions and u_i denotes the idiosyncratic error term. All remaining elements are defined as in (2.1).

Taking conditional expectations on both sides of the structural equation (2.6) and invoking the conditional moment restrictions above, it follows that the conditional mean of (2.6) no longer coincides with the structural object of interest.

In particular, $E[Y_i|X_{1i}, X_{2i}, W_i^c, Z_i, u_i] \neq m(X_{1i}, W_i^c) + \beta^\top X_{2i}$,

$$\begin{aligned} E[Y_i|X_{1i}, X_{2i}, W_i^c, Z_i, u_i] &= m(X_{1i}, W_i^c) + \beta^\top X_{2i} + E[\epsilon_i|X_{1i}, X_{2i}, W_i^c, Z_i, u_i] \\ &= m(X_{1i}, W_i^c) + \beta^\top X_{2i} + E[\epsilon_i|u_i], \end{aligned} \quad (2.7)$$

because $E[\epsilon_i|X_{1i}, X_{2i}, W_i^c, Z_i, u_i] = E[\epsilon_i|X_{1i}, X_{2i}, Z_i, u_i] = E[\epsilon_i|u_i]$.

In this chapter, we adopt the CF approach, as it offers important advantages for the treatment of nonlinear models. In particular, it is computationally less demanding than maximum likelihood methods and typically requires fewer distributional assumptions.

The CF method addresses endogeneity by incorporating first-stage residuals into the structural equation, thereby capturing the dependence structure between the endogenous regressor and the structural disturbance. Although both IV and CF methods rely on exclusion restrictions and valid instruments, the CF framework may yield efficiency gains in nonlinear settings by exploiting the triangular structure assumption in 2.6. Such gains, however, are obtained at the cost of requiring correct specification of the first-stage equation: if the first-stage is misspecified, the resulting CF estimator will generally be inconsistent, even in situations where standard IV procedures remain consistent (Wooldridge 2010). Newey, Powell, & Vella (1999) extend this reasoning to the nonparametric setting, allowing the CF to enter the outcome equation flexibly and without imposing restrictive parametric assumptions on the structural equation.

Following the logic of the CF approach, we augment the structural equation in (2.6) by introducing an additional smooth function that captures the relationship between the endogenous variables and the structural error term. The resulting regression model is given by

$$Y_i = m(X_{1i}, W_i^c) + \beta^\top X_{2i} + \theta(u_i) + v_i, \quad (2.8)$$

where $\theta(u_i) = E[\epsilon_i|u_i]$ is an unknown smooth function, with $\theta(\cdot) : \mathbb{R}^{q_w} \rightarrow \mathbb{R}$, and $v_i = Y_i - E[Y_i|X_{1i}, X_{2i}, W_{1i}^c, Z_i, u_i]$ denotes a purely random disturbance. Direct estimation of this model is infeasible because u_i is unobservable and must therefore be estimated in a preliminary step.

2.2.2 Estimation Procedure

To obtain feasible and consistent estimates of equation (2.8), we propose a three-stage estimation procedure which is described in detail below. This methodology addresses the endogeneity problem sequentially by estimating the relevant components in stages, while preserving robustness in the presence of nonlinearities.

First Stage

In the initial stage, we estimate the endogenous component of the model. Specifically, when all excluded instruments in Z_i are continuous, that is, when $Z_i \equiv Z_i^c$, the first-stage regression takes the form

$$W_i^c = g(Z_i^c) + u_i. \quad (2.9)$$

A consistent estimator of $g(\cdot)$ may be obtained using a local constant (linear) kernel approach. In particular, the Nadaraya-Watson estimator provides a natural and widely used nonparametric approach, owing to its simplicity and its capacity to approximate complex functional relationships with considerable flexibility.

However, the local constant estimator may suffer from a substantial bias when estimating the regression function near the boundary of the support, leading to inaccurate results in regions with limited data density (Fan & Gijbels 1992). To overcome this issue, we adopt the local linear approach proposed in Stone (1977) and Cleveland (1979), since this method retains many of the desirable properties of the local constant approach (such as identical variance) while offering supe-

rior boundary correction capabilities. More precisely, unlike the local constant method, whose bias depends on the underlying density function $f(z)$, the local linear estimator corrects for boundary effects without such dependence, making it a more reliable technique in finite samples (Fan 1992; Fan & Gijbels 1992).

To derive a more accurate local approximation to $g(z) \equiv g(z^c)$, let z^c denote a point in the support of Z_i^c . A first-order Taylor expansion of $g(\cdot)$ around z^c yields the following linear approximation

$$W_i^c \approx g(z^c) + \sum_{s=1}^{q_z} (Z_s^c - z^c)^\top \nabla_{g_s}(z^c) + u_i, \quad (2.10)$$

where $\nabla_{g_s}(z^c) = \frac{\partial g(z^c)}{\partial Z_s^c}$ denotes the $(q_z \times 1)$ gradient vector containing the first-order partial derivatives of $g(\cdot)$ with respect to Z_s^c .

In practice, observations located far from z^c provide relatively poor approximations to $g(z^c)$, since the discrepancy between their associated function values and the true function value at z^c is typically larger. Treating all observations symmetrically would therefore increase the risk of estimation bias. To alleviate this problem, kernel weighting is employed to emphasize local information: observations in the neighborhood of z^c receive greater weight, whereas the contribution of more distant observations is progressively downweighted. In this way, estimation is driven primarily by those data points that offer the most reliable local approximation to $g(z^c)$. Hence, the local linear estimator is obtained by solving the following minimization problem

$$\min_{a,b} \sum_{i=1}^N \left(W_i^c - a - (Z_i^c - z^c)^\top b \right)^2 K_{h_1}(Z_i^c - z^c), \quad (2.11)$$

where $K_{h_1}(Z_i^c - z^c) = \frac{1}{h_1^{q_z}} k\left(\frac{Z_{i1}^c - z_1^c}{h_1}\right) \times \dots \times k\left(\frac{Z_{iq_z}^c - z_{q_z}^c}{h_1}\right) = \frac{1}{h_1^{q_z}} \prod_{s=1}^q k\left(\frac{Z_{is}^c - z_s^c}{h_1}\right)$ is a kernel function with first-stage bandwidth parameter h_1 , a represents the local approximation of $g(z^c)$, and b is the corresponding value for the gradient function $\nabla_g(z^c)$. Let $\widehat{a} = \widehat{g}(z^c)$ and $\widehat{b} = \widehat{\nabla}_g(z^c)$ denote the solutions to (2.11). The resulting nonpara-

metric estimator of $g(z^c)$ can then be expressed as

$$\widehat{g}(z^c; h_1) = e_1^\top \left[\sum_{i=1}^N K_{h_1}(Z_i^c, z^c) \begin{pmatrix} 1, & (Z_i^c - z^c)^\top \\ (Z_i^c - z^c), & (Z_i^c - z^c)(Z_i^c - z^c)^\top \end{pmatrix} \right]^{-1} \sum_{i=1}^N K_{h_1}(Z_i^c, z^c) \begin{pmatrix} 1 \\ (Z_i^c - z^c) \end{pmatrix} W_i^c, \quad (2.12)$$

where $e_1 = (1, 0_{q_z})$ is the $(q_z + 1) \times 1$ vector whose first entry is equal to one and whose remaining entries are zero.

When Z_i contains continuous and discrete variables, $Z_i \equiv (Z_i^c, Z_i^d)$, a different estimation strategy is required to obtain a consistent estimator of $g(\cdot)$, since the vector of instruments Z_i contains variables of different natures. Nonparametric estimation with mixed data types raises important methodological difficulties. Although the nonparametric procedures discussed above are well-suited to continuous regressors, their direct application to categorical variables is neither straightforward nor especially effective.

The conventional approach to handling discrete regressors consists in partitioning the sample into subsets, or "cells", corresponding to every possible combination of categories, and then applying nonparametric methods within each cell (Aitchison & Aitken 1976). This strategy, however, suffers from serious limitations. First, when the number of possible category combinations is large, as occurs with multiple categorical variables taking several values, the number of cells increases exponentially. This well-known curse of dimensionality leads to many cells containing few, if any, observations, thereby producing imprecise and unreliable estimates (Scott 1992; Pagan & Ullah 1999). Second, such sample fragmentation reduces estimator efficiency, inflates variance, and hinders the identification of meaningful relationships involving the continuous covariates (Li & J. S. Racine 2007). To overcome these limitations, we follow J. Racine & Li (2004b) who propose a nonparametric method extending the kernel estimator of Aitchison & Aitken (1976) for mixed data to the multivariate setting. This approach

avoids partitioning the sample into cells, thereby preserving the computational tractability of kernel methods while delivering more stable and accurate estimates in the presence of mixed regressors.

In this case, the first-stage regression model is

$$W_i^c = g(Z_i^c, Z_{1i}^d) + u_i \quad (2.13)$$

where $g(\cdot)$ is an unknown function, Z_{1i}^c is a $(q_z^c \times 1)$ vector of continuous variables, and Z_i^d is a $(q_z^d \times 1)$ vector of unordered discrete variable. The extension to the case in which the discrete variables possess a natural ordering lies beyond the scope of this study.

Using the mixed-data product kernel introduced at the beginning of the chapter, and denoting $z \equiv (z^c, z^d) \in \mathbb{R}_z^q$, where $q = (q_z^c + q_z^d)$, we define the following product of kernels:

$$\begin{aligned} \mathcal{W}_\phi(z^c, z^d) &= \frac{1}{h_1^{q_z^c}} \prod_{s=1}^{q_z^c} k\left(\frac{Z_{is}^c - z_s^c}{h_1}\right) \prod_{r=1}^{q_z^d} \ell(Z_{ir}^d, z_r^d, \lambda_1) \\ &= K_{h_1}(Z_i^c, z^c) L(Z_i^d, z^d, \lambda_1). \end{aligned} \quad (2.14)$$

Note that $\lambda_1 \in [0, 1]$. When $\lambda_1 = 0$, the discrete kernel reduces to an indicator function, whereas $\lambda_1 = 1$ yields a uniform weighting scheme for $L(\cdot)$. Moreover, $L(Z_i^d, z^d, \lambda_1)$ does not, in general, sum to one whenever $\lambda_1 \neq 0$. As noted by J. Racine & Li (2004b), however, this feature does not affect the resulting nonparametric estimator, since the kernel appears in both the numerator and denominator. Consequently, the kernel function may be multiplied by any nonzero constant without altering the estimator of $g(\cdot)$.

Following J. Racine & Li (2004b), a Nadaraya-Watson estimator could be constructed using the kernel weight defined above. Nevertheless, in view of the superior properties of local linear estimators, we instead adopt the local linear least squares approach of J. Racine & Li (2004a). Let $\nabla g_1(z^c, z^d)$ denote the partial

derivative of $g(z) = g(z^c, z^d)$ with respect to Z_i^c , that is, (i.e., $\nabla g_1(z^c, z^d) = \frac{\partial g(z^c, z^d)}{\partial Z_i^c}$).

The, for any given z^c and $\widetilde{z}^d$ in a neighborhood of z , a first-order Taylor expansion implies that

$$g(\widetilde{z}) \approx g(z^c, z^d) + \nabla_{g_1}(z^c, z^d)(Z_i^c - z^c). \quad (2.15)$$

To estimate $g(\widetilde{z})$ and $\nabla_g(z^c, z^d)$, we choose a and b to minimize

$$\sum_{i=1}^N \left[W_i^c - a - (Z_i^c - z^c)^\top b \right]^2 \mathcal{W}_\phi(Z_i, z). \quad (2.16)$$

Let $\{\widehat{a}, \widehat{b}\}$ denote the resulting local linear estimator. The estimator of the functional coefficient is then given by

$$\widehat{g}(z, \phi) = e_1^\top \left[\sum_{i=1}^N \mathcal{W}_\phi(Z_i, z) \begin{pmatrix} 1, & (Z_i^c - z^c)^\top \\ (Z_i^c - z^c), & (Z_i^c - z^c)(Z_i^c - z^c)^\top \end{pmatrix} \right]^{-1} \sum_{i=1}^N \mathcal{W}_\phi(Z_i, z) \begin{pmatrix} 1 \\ (Z_i^c - z^c) \end{pmatrix} W_{1i}^c, \quad (2.17)$$

where $e_1 = (1, 0_{q_z^c})$ is the $(q+1) \times 1$ vector whose first component is equal to one and whose remaining components are zero.

Second Stage

To control for endogeneity in (2.6), the second stage uses the nonparametric residuals obtained from the first stage, namely,

$$\widehat{u}_i = W_i^c - \widehat{g}(Z_i),$$

where $\widehat{g}(Z_i) \equiv \widehat{g}(Z_i; h_1)$ when the excluded instruments are continuous, $\widehat{g}(Z_i) \equiv \widehat{g}(Z_i; \lambda_1)$ when they are discrete, and $\widehat{g}(Z_i) \equiv \widehat{g}(Z_i; h_1, \lambda_1)$ when they are mixed. Let $m_0(X_{1i}, W_i^c, u_i) = m(X_{1i}, W_i^c) + \theta(u_i)$. The regression model to estimate may then be written as

$$Y_i = m_0(X_{1i}, W_i^c, \widehat{u}_i) + \beta^\top X_{2i} + v_i^*, \quad (2.18)$$

where $v_i^* = \epsilon_i + [m_0(X_{1i}, W_i^c, u_i) - m_0(X_{1i}, W_i^c, \widehat{u}_i)]$ is the resulting error term.

To eliminate the nonparametric component from the preceding partially linear model, we follow the idea in Robinson (1988) and take the conditional expectation on both sides of (2.18), obtaining

$$E[Y_i | X_{1i}, W_i^c, \widehat{u}_i] = m_0(X_{1i}, W_i^c, \widehat{u}_i) + \beta^\top E[X_{2i} | X_{1i}, W_i^c, \widehat{u}_i] + E[v_i^* | X_{1i}, W_i^c, \widehat{u}_i]. \quad (2.19)$$

Subtracting (2.19) from (2.18) yields

$$Y_i - E[Y_i | X_{1i}, W_i^c, \widehat{u}_i] = \beta^\top \{X_{2i} - E[X_{2i} | X_{1i}, W_i^c, \widehat{u}_i]\} + v_i^* \quad (2.20)$$

since $E[v_i^* | X_{1i}, W_i^c, \widehat{u}_i] = 0$ by construction. Feasible estimation of β requires estimation of the relevant conditional expectations. To this end, we employ the local-constant approach of Robinson (1988), although a local-linear approach could also be used (Delgado & Parmeter 2014).

Let $\widehat{Y}_i = Y_i - \widehat{E}[Y_i | X_{1i}, W_i^c, \widehat{u}_i]$ and $\widehat{X}_{2i} = X_{2i} - \widehat{E}[X_{2i} | X_{1i}, W_i^c, \widehat{u}_i]$, where $\widehat{E}[Y_i | X_{1i}, W_i^c, \widehat{u}_i]$ and $\widehat{E}[X_{2i} | X_{1i}, W_i^c, \widehat{u}_i]$ denote the corresponding Nadaraya-Watson estimators of $E[Y_i | X_{1i}, W_i^c, \widehat{u}_i]$ and $E[X_{2i} | X_{1i}, W_i^c, \widehat{u}_i]$, respectively. An OLS regression of $\widehat{Y}_i$ on

$\widehat{\widehat{X}}_{2i}$ then yields

$$\widehat{\beta} = \left(\sum_{i=1}^N \widehat{\widehat{X}}_{2i} \widehat{\widehat{X}}_{2i}^\top \widehat{f}^2(X_{1i}, W_{1i}^c, \widehat{u}_i) \right)^{-1} \sum_{i=1}^N \widehat{\widehat{X}}_{2i} \widehat{\widehat{Y}}_i \widehat{f}^2(X_{1i}, W_{1i}^c, \widehat{u}_i), \quad (2.21)$$

where $\widehat{f}(X_{1i}, W_{1i}^c, \widehat{u}_i)$ is an estimator of the density $f(X_{1i}, W_{1i}^c, \widehat{u}_i)$, introduced to avoid the random-denominator problem associated with the Nadaraya-Watson estimator (Powell et al. 1989). Its form is given by

$$\widehat{f}(X_{1i}, W_{1i}^c, \widehat{u}_i) = \frac{1}{N h_2^{q_{x_1} + q_{w_1} + 1}} \sum_{i=1}^N k\left(\frac{X_{1i} - x_1}{h_{2x_1}}\right) k\left(\frac{W_{1i}^c - w^c}{h_{2w^c}}\right) k\left(\frac{\widehat{u}_i - \widehat{u}}{h_{2u}}\right).$$

Third Stage

We now turn to estimation of the unknown smooth function $m(\cdot)$. Using the second stage results, the partially linear model may be rewritten as a purely non-parametric regression. Specifically, defining $Y_i^* = Y_i - \widehat{\beta}^\top X_{2i}$, we obtain

$$Y_i^* = m_0(X_{1i}, W_{1i}^c, \widehat{u}_i) + v_i^*, \quad (2.22)$$

where $v_i^* = v_i^* - (\widehat{\beta} - \beta)^\top X_{2i}$.

A nonparametric estimator may then be used to recover $m_0(\cdot)$. Let $\chi_i = (X_{1i}, W_{1i}^c, \widehat{u}_i)$.

Employing once again a local linear approach, we define the estimator

$$\begin{aligned} \widehat{m}_0(x_1, w^c, \widehat{u}) &= e_1^\top \left[\sum_{i=1}^N K_{h_3}(\chi_i, \chi_0) \begin{pmatrix} 1 & (\chi_i - \chi_0)^\top \\ (\chi_i - \chi_0) & (\chi_i - \chi_0)(\chi_i - \chi_0)^\top \end{pmatrix} \right]^{-1} \\ &\quad \times \sum_{i=1}^N K_{h_3}(\chi_i, \chi_0) \begin{pmatrix} 1 \\ (\chi_i - \chi_0) \end{pmatrix} Y_i^*, \end{aligned} \quad (2.23)$$

where χ_0 is an interior point in the support of χ_i , and h_3 denotes the third-stage bandwidth.

Finally, we estimate the structural function $m(\cdot)$ by means of the marginal integration approach independently introduced by Tjøstheim & Auestad (1994),

Newey (1994), and Linton & Nielsen (1995) obtaining

$$\widehat{m}(x_1, w^c) = \int \widehat{m}_0(x_1, w^c, \widehat{u}) \omega(\widehat{u}) d\widehat{u}, \quad (2.24)$$

where $\omega(\cdot)$ is a predetermined positive weighting function satisfying $\int_{\mathbb{W}} \omega(u) du = 1$, that is, $\omega(\cdot)$ is a density function on the compact support $\mathbb{W}$ of u_i .

In principle, the marginal integration in (2.24) may be implemented numerically using quadrature methods such as Simpson's or the Trapezoidal rules. Let $\mathcal{S}$ denote the number of evaluation points along each "nuisance dimension", namely, each dimension of $\widehat{u}_i$. Then, for every point in the dimension of interest, estimation requires the computation of $\mathcal{S}^{(q_{x_1} + q_w)}$ weighted least squares problems, each involving $O(N)$ operations. This may pose a practical computational burden and therefore requires numerical precision against computational cost.

To address this issue, a computationally less demanding alternative is available; see Qian & Wang (2012) & Delgado & Parmeter (2014). The basic idea is to generate independent and identically distributed (*i.i.d.*) samples $(\widehat{u}_\ell^*, \ell = 1, \dots, N)$ from the distribution $\omega(\cdot)$ and construct $\widehat{m}_{MC}(x_1, w^c) = \frac{1}{N} \sum_{\ell=1}^N \widehat{m}_0(x_1, w^c, \widehat{u}_\ell^*)$. For sufficiently large N , $\widehat{m}_{MC}(\cdot)$ provides an accurate approximation to $\widehat{m}(\cdot)$. A particularly convenient choice is to let $\omega(\cdot)$ coincide with the density of $\widehat{u}_i$, in which case the sample analogue becomes

$$\widehat{m}(x_1, w^c) = \frac{1}{N} \sum_{i=1}^N \widehat{m}_0(x_1, w^c, \widehat{u}_i) \quad (2.25)$$

which is asymptotically equivalent to (2.23) provided that $E(\epsilon_i) = 0$.

It is worth noting that this final step requires estimating $\widehat{m}_0(x_1, w^c, \cdot)$ at each value of $\widehat{u}_i$ and then averaging across the sample. However, reliable estimates of $m(\cdot)$ do not necessarily require integration over the full sample. As noted by Henderson & Parmeter (2015), using approximately 25% of the overall sample yields estimates with bias and variance very similar to those obtained using the complete

sample.

2.2.3 Bandwidth Selection Procedure

As noted above, the choice of the bandwidth parameters h and λ is a crucial component of kernel regression, since it governs the trade-off between bias and variance. The bandwidth determines the finite-sample behavior of the estimator, whereas the kernel function shapes its smoothness and differentiability properties. In general, the optimal bandwidth is selected so as to minimize a prescribed loss criterion. In finite samples, the mean squared error (MSE) is commonly employed, as it balances estimation accuracy against model complexity. In contrast, asymptotic approaches typically minimize the asymptotic mean integrated squared error (AMISE), while fully data-driven procedures often target predictive performance.

A variety of bandwidth selection methods have been proposed, including Rule-of-Thumb (RoT) procedures, plug-in methods, cross-validation (CV), and bootstrap-based techniques. Among these, CV is widely used, particularly in local linear kernel estimation, and remains especially suitable in the presence of discrete regressors (P. Hall, J. Racine, et al. 2004).

Within triangular simultaneous equation models such as (2.6), an important theoretical result is that the asymptotic distribution of the local polynomial estimator is unaffected by the choice of the first-stage bandwidth; see Su & Ullah (2008) for further details. Moreover, as emphasized by Henderson & Parmeter (2015), selecting an optimal first-stage bandwidth does not necessarily guarantee optimal performance of the resulting nonparametric estimator. In response to this issue, Delgado & Parmeter (2014) advocate least-squares cross-validation (LSCV) in both stages, whereas Martins-Filho & Yao (2012) rely on a simple RoT procedure.

For the sake of simplicity, in line with Martins-Filho & Yao (2012), we adopt the RoT method in all three stages, as it provides a rapid and straightforward means of selecting bandwidths. In the nonparametric framework, however, the formal theory underlying RoT bandwidth choice is well developed only for continuous covariates, not for discrete ones. Accordingly, for continuous variables the rule typically scales the bandwidth by the standard deviation of the regressor, thereby reflecting its dispersion and scale. No analogous theoretical benchmark exists for discrete variables. In such cases, one must rely on practical conventions. One such strategy, proposed by Henderson & Parmeter (2015), is to employ a pilot bandwidth based primarily on sample size, which becomes the main informative component in the absence of a meaningful measure of distributional spread. More specifically, Technical Appendix 8 in Henderson & Parmeter (2015) shows that the optimal bandwidth for each type of regression is proportional to sample size and may be written as

$$h \sim N^{-\frac{1}{4+q_c}} \quad \text{and} \quad \lambda \sim N^{-\frac{2}{4+q_c}},$$

where q_c represents the number of continuous regressors under consideration. For continuous covariates, the bandwidth is further rescaled by the variability of the regressor, yielding

$$h = c\sigma_x N^{-\frac{1}{4+q_c}},$$

where c is a scale constant and σ_x denotes the standard deviation of the covariate. For discrete covariates, in the absence of a formal rule, we follow the practical recommendation in Henderson & Parmeter (2015) and adopt the pilot bandwidth

$$\lambda = cN^{-\frac{2}{4+q_c}}.$$

2.3 Monte Carlo Simulations

To evaluate the performance of the estimators proposed in this chapter, we conduct a Monte Carlo simulation based on the following data-generating processes (DGPs),

$$\text{DGP1: } Y_i = \frac{\exp(X_{1i} + W_{1i}^c)}{1 + 3 \exp(X_{1i} + W_{1i}^c)} + \beta X_{2i} + \epsilon_i, \quad (2.26)$$

$$\text{DGP2: } Y_i = \frac{-1}{(2X_{1i} + \sin(2W_{1i}^c))} + \beta X_{2i} + \epsilon_i, \quad (2.27)$$

where $\beta = 1$ and the covariates are generated as random variables such that $(X_{1i}, X_{2i}) \sim i.i.d. \mathcal{U}(-5, 5)$. To induce correlation between W_{1i}^c and the disturbance term, we specify

$$W_{1i}^c = c_0 Z_i + u_i \quad (2.28)$$

where c_0 is a constant governing the instrument's strength, with $c_0 = 1$ corresponding to strong instruments and $0 < c_0 < 1$ to weaker instruments, and $Z_i \sim i.i.d. \mathcal{U}(-5, 5)$. In addition, we assume that ϵ_i and u_i are jointly normally distributed according to

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \sim i.i.d. \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix} \right), \quad (2.29)$$

where $\theta = (0, 0.3, 0.6)$ governs the degree of endogeneity, ranging from the exogenous case ($\theta = 0$) to settings with low and high levels of endogeneity.

The number of Monte Carlo replications is set to 1,000. We employ the Epanechnikov kernel function $k(u) = 0.75(1 - u^2)1\{|u| \leq 1\}$, and consider sample sizes $N \in \{100, 200, 400\}$. To evaluate the performance of the estimator proposed for β in (2.21), we report bias, standard deviation (Sd), and Root Mean Squared Error (RMSE). For the nonparametric estimator of $m(\cdot)$ in (2.23), performance is

assessed using the RMSE. These measures are computed as

$$RMSE_{\beta} = \left[\frac{1}{R} \sum_{r=1}^R (\widehat{\beta}_r - \beta)^2 \right]^{\frac{1}{2}} \quad (2.30)$$

$$RMSE_m = \left[\frac{1}{N} \sum_{i=1}^N (\widehat{m}_r(x_{1i}, w_i^c) - m_r(x_{1i}, w_i^c))^2 \right]^{\frac{1}{2}}, \quad (2.31)$$

where the subscript r denotes the r -th replication out of a total of R .

To ensure robustness, the analysis is restricted to the 10%-90% quantile range of the simulation outcomes, thereby excluding extreme values, as suggested by Martins-Filho & Yao (2012) and Delgado & Parmeter (2014).

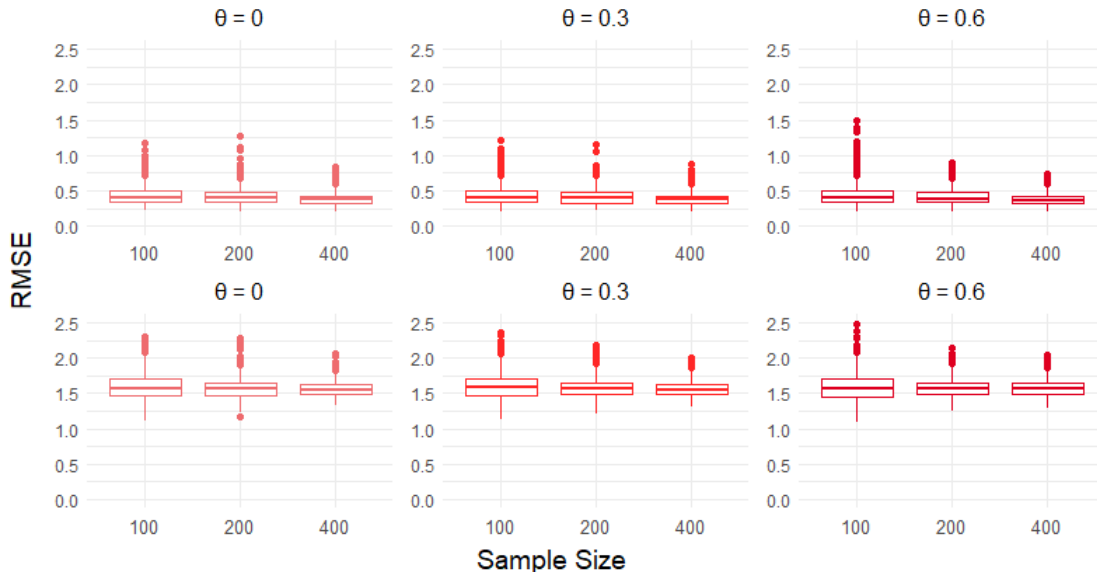
Table 2.1 reports the Monte Carlo results for $\widehat{\beta}$, confirming that the proposed three-stage estimator exhibits strong consistency properties under both DGPs. For all values of θ , the standard deviation and RMSE of $\widehat{\beta}$ decline systematically as the sample size increases from $N = 100$ to $N = 400$. This pattern persists even under a high degree of endogeneity ($\theta = 0.6$), suggesting that the estimator remains stable and convergent in increasingly endogenous environments. Overall, the results indicate that the estimator displays reliable finite-sample performance, with no evidence of instability as the level of endogeneity intensifies.

Figure 2.1 illustrates the distribution of the RMSE of the nonparametric estimator across different sample sizes and degrees of endogeneity (θ), using boxplots based on repeated Monte Carlo replications. The results indicate a consistent improvement in performance as the sample size increases, reflected in a systematic reduction in RMSE. Importantly, even under moderate and high levels of endogeneity ($\theta = 0.3$ and 0.6), the estimation error remains well behaved and bounded, with no evidence of divergence, in accordance with the theoretical consistency of the estimator. Furthermore, for a given sample size, the RMSE exhibits remarkable stability across different levels of endogeneity.

Table 2.1: Finite sample performance of $\widehat{\beta}$

θ	N	DGP1			DGP2		
		<i>Bias</i>	<i>Sd</i>	$RMSE_{\beta}$	<i>Bias</i>	<i>Sd</i>	$RMSE_{\beta}$
0	100	-0.0015	0.0544	0.0544	0.0030	0.0677	0.0678
	200	-0.0034	0.0416	0.0418	-0.0029	0.0495	0.0496
	400	-0.0014	0.0284	0.0285	-0.0029	0.0356	0.0357
0.3	100	-0.0058	0.0523	0.0526	-0.0075	0.0645	0.0650
	200	-0.0106	0.0412	0.0426	-0.0096	0.0489	0.0498
	400	-0.0145	0.0286	0.032	-0.0155	0.0360	0.0392
0.6	100	-0.0118	0.0532	0.0545	-0.0127	0.0656	0.0668
	200	-0.0206	0.0421	0.0469	-0.0198	0.0495	0.0533
	400	-0.0280	0.0292	0.0405	-0.0293	0.0366	0.0469

Note: The table reports average Bias, standard deviation (Sd), and RMSE of $\widehat{\beta}$ over 1,000 replications for the three-stage estimator under DGP1 and DGP2.

Figure 2.1: Finite sample performance of $\widehat{m}(\cdot)$ 

Note: Boxplot of the RMSE values of the nonparametric estimator of $m(\cdot)$ in 1,000 replications.

2.4 Conclusions

This chapter revisits the control function approach to endogeneity within a semi-parametric IVs framework for partially linear models. In contrast to standard settings that restrict instruments to be continuous, the proposed estimation procedure accommodates both continuous and/or discrete instruments through the use of generalized product kernels (J. Racine & Li 2004b). This approach enables the treatment of unordered discrete variables in a unified framework, avoiding the need for ad hoc discretization or separate estimation across subgroups.

The proposed three-stage estimator proceeds by first constructing residuals from a nonparametric regression of the endogenous continuous regressor on the full set of instruments. In the second stage, Robinson (1988)'s technique is employed to estimate the parametric component. Finally, the nonparametric structural function is recovered through marginal integration. Identification is achieved via the CF condition: conditional on the first-stage residual, the endogenous regressor is rendered exogenous in the second-stage regression.

The Monte Carlo results presented in this chapter are in line with the theoretical properties of the estimator. In particular, the parametric component achieves $\sqrt{N}$ -consistency and asymptotic normality, while the nonparametric component converges at the standard kernel rate.

Chapter 3

Dummy Endogenous Variables in Partially Linear Regression Models

Abstract

We propose a novel estimator for partially linear models in which endogenous dummy variables enter the parametric component. The proposed approach integrates local likelihood methods, instrumental variable techniques, and local linear regression within a unified estimation framework. The resulting procedure delivers $\sqrt{N}$ -consistent estimates for the parametric coefficients, while the nonparametric component attains the optimal convergence rate. Monte Carlo simulations provide evidence of strong finite-sample performance across different levels of endogeneity. Finally, the practical relevance of the methodology is illustrated through an empirical application to the estimation of a knowledge capital production function.

3.1 Introduction

Since the seminal contribution of J. J. Heckman (1978), substantial attention has been devoted to simultaneous two-equation systems in which latent continuous variables generate observed binary outcomes. A particularly relevant case is the so-called hybrid model, in which one dependent variable is observed while the other remains latent (Maddala 1983). The specification may also include the observed binary counterpart of the latent variable as an additional regressor, thereby allowing for a possible structural shift.

The dummy endogenous variable model (DEVIM) constitutes a special case of Heckman's hybrid model, in which the equation for the continuous dependent variable includes the observed endogenous dummy variable, rather than its latent counterpart, as a regressor. This hybrid specification was first proposed by Landes (1968) in their analysis of the effects of fair-employment legislation on the status of black workers, and later employed by Edwards (1978) to estimate the impact of compulsory school enrollment laws. Maddala (1983) provides a detailed discussion of this model and shows how it can be estimated by full-system maximum likelihood, an approach often referred to as the Heckman-Maddala estimator. Subsequent contributions (J. D. Angrist 2001; Wooldridge 2003; Wooldridge 2010) explored instrumental variable (IV) methods as an alternative, emphasizing the trade-off between the efficiency full-system maximum-likelihood and the robustness of single-equation IV estimators.

A comprehensive overview of the econometric issues associated with models containing endogenous dummy regressors is provided by Vella (1998). Other surveys have highlighted their empirical relevance, particularly in fields such as education, labor economics, and innovation, where key explanatory variables frequently take binary form (R. Blundell & Dias 2009; Mairesse & Mohnen 2010).

Despite their simplicity and straightforward interpretation, however, DEVMs are vulnerable to misspecification bias when the assumed parametric structure is incorrect.

Partially linear models (PLMs) offer an appealing alternative. Their modular structure allows certain covariates to enter nonparametrically, thereby preserving flexibility, while retaining a parametric specification for others and thus avoiding the curse of dimensionality that may arise in fully nonparametric models (Robinson 1988). Typical examples include log-wage equations in which experience is modeled nonparametrically while other characteristics enter linearly (Hsu et al. 2022); electricity demand equations in which consumption is represented as a smooth function of temperature and as a linear function of price and income (Engle et al. 1986); and labor supply models in which lagged hours and real wages are specified nonparametrically while the remaining controls enter parametrically (Gao et al. 2015).

Over recent decades, the econometric literature has made substantial progress in developing consistent estimation methods for semiparametric models with continuous endogenous regressors (Newey & Powell 2003; R. W. Blundell & Powell 2004). A considerable body of work has examined endogeneity in PLMs (Newey, Powell, & Vella 1999; Pinkse 2000; Newey & Powell 2003; Su & Ullah 2008; Martins-Filho & Yao 2012; Delgado & Parmeter 2014); however, none of these approaches explicitly accommodates endogenous dummy covariates. Das (2005) introduces the endogenous dummy variable in a fully nonparametric framework using series methods, allowing its effect to vary nonparametrically with continuous covariates, in a manner similar to Cardot & Musolesi (2020) in the context of panel data. This degree of flexibility, however, comes at the cost of the curse of dimensionality, reduced interpretability, and substantial computational difficulties, including slow convergence, high sampling variability, and pronounced

sensitivity to tuning parameters.

This chapter proposes a new estimation procedure for PLMs with endogenous dummy variables. The proposed specification extends the DEVM framework by allowing exogenous covariates to enter nonparametrically, while the endogenous dummy variable induces a homogeneous structural shift. In this way, the approach combines the flexibility of semiparametric models with the interpretability of conventional parametric methods, while mitigating the misspecification bias associated with overly restrictive functional form assumptions.

The structure of this chapter proceeds as follows. Section 3.2 formally introduces the econometric specification under analysis and develops the consistent estimation of PLM with endogenous dummy variables in the parametric component. Section 3.3 presents Monte Carlo simulations designed to assess the finite-sample performance of the proposed estimator. Section 3.4 applies the methodology to the estimation of a knowledge production function. Finally, Section 3.5 concludes.

3.2 Econometric Model and Estimation Procedure

We consider a partially linear structural equation of the form

$$Y_i = m(X_{1i}) + X_{2i}^\top \beta + W_i^d \gamma + \epsilon_i, \quad (3.1)$$

where $X_{1i} \in \mathbb{R}^{q_{x1}}$ and $X_{2i} \in \mathbb{R}^{q_{x2}}$ are observed exogenous covariates, and $W_i^d \in \{0, 1\}^{q_w}$ is a dummy endogenous variable. The function $m(\cdot) : \mathbb{R}^{q_{x1}} \rightarrow \mathbb{R}$ is unknown, while $\beta \in \mathbb{R}^{q_{x2}}$ and $\gamma \in \mathbb{R}^{q_w}$ are finite-dimensional parameters to be estimated. The term ϵ_i denotes an unobserved disturbance.

The dummy endogenous variable is generated according to

$$W_i^d = \mathbb{1}\{g(Z_i) + u_i > 0\}, \quad (3.2)$$

where Z_i is a vector of exogenous variables satisfying $E[\epsilon_i | Z_i] = 0$, and u_i is a reduced-form error term such that $E[u_i | Z_i] = 0$.

Endogeneity of W_i^d arises because ϵ_i and u_i are correlated, so that $E[\epsilon_i | u_i] \neq 0$. Since W_i^d is a function of u_i through (3.2), this dependence implies $E[\epsilon_i | W_i^d] \neq 0$, whereas the exclusion restriction ensures that $E[\epsilon_i | Z_i] = 0$. As a consequence, the conditional mean of Y_i does not directly identify the structural parameters. Taking conditional expectations on both sides of (3.1) yields

$$E[Y_i | X_{1i}, X_{2i}, W_i^d, Z_i, u_i] = m(X_{1i}) + X_{2i}^\top \beta + W_i^d \gamma + E[\epsilon_i | u_i], \quad (3.3)$$

where the final term follows from the assumption that $E[\epsilon_i | X_{1i}, X_{2i}, Z_i, u_i] = E[\epsilon_i | u_i]$. That is, once one conditions on u_i , the exogenous covariates X_{1i} , X_{2i} , and the excluded instruments Z_i provide no additional information about ϵ_i ; moreover, W_i^d drops out because it is itself a function of (Z_i, u_i) .

A widely used strategy for addressing this problem is the CF approach, which seeks to purge W_i^d of its endogenous component through an auxiliary regression (J. J. Heckman 1979; Rivers & Vuong 1988). In the case of a continuous endogenous regressor, this procedure is relatively straightforward: the first-stage OLS residual $\hat{u}_i$ consistently estimates u_i and therefore provides a valid correction term without requiring strong distributional assumptions (R. Blundell & Powell 2003). In the binary case, however, the approach is subject to several important limitations that substantially reduce its applicability and reliability:

- Unobserved residual problem: Since W_i^d is an indicator of an unobserved latent variable, namely $W_i^{d*} = g(Z_i) + u_i$, the disturbance u_i cannot be recovered exactly. Its distribution must instead be explicitly modeled to construct any usable proxy, rendering all existing CF implementations inherently parametric. In practice, CF corrections rely on approximations to u_i , such as the inverse Mills ratio or generalized residuals (J. J. Heckman

1979; Gourieroux et al. 1987), which are valid only under strong distributional assumptions, most notably joint normality of the error terms. Consequently, any CF correction for W_i^d is sensitive to misspecification of the first-stage distribution in a way that has no analogue in the continuous case (Wooldridge 2015).

- Functional form dependence: The validity of the CF correction depends critically on the correct specification of the first-stage model. If the functional form, for example the index structure or distribution of u_i , is misspecified, the correction fails to eliminate endogeneity, thereby generating biased estimates.
- Incomplete endogeneity correction: In the presence of complex error structures, such as nonlinear interactions or heteroskedasticity, CF corrections based on a single residual term may be insufficient to restore orthogonality between the regressors and ϵ_i . This limitation is particularly problematic in semiparametric models, where non-additive interactions between observables and unobservables are often more plausible.

These concerns are further compounded by the forbidden regression problem (Wooldridge 2010), which arises when fitted values from a nonlinear first-stage model, such as probit or logit, are inserted directly as regressors in the second-stage equation. This practice violates the orthogonality conditions required for valid IV estimation and therefore leads to inconsistent estimates.

The estimation procedure proposed in this chapter directly addresses these limitations. Rather than relying on CF corrections or substituting nonlinear fitted values for the endogenous regressor, we adopt an IV strategy in which the conditional probability $p_i = P(W_i^d = 1 \mid Z_i)$ is used as an instrument for W_i^d in the structural equation (3.1). By construction, p_i inherits exogeneity from Z_i , so that

$E[\epsilon_i \mid p_i] = 0$. This strategy preserves consistency without imposing distributional assumptions on ϵ_i or restricting the form of its interaction with observables (Wooldridge 2015).

Before turning to the estimation procedure, it is useful to introduce the kernel functions employed to smooth over both continuous and categorical variables. Let $Z_i = (Z_i^c, Z_i^d)$ be a $(q_z \times 1)$ vector of IV, where Z_i^c denotes the first q^c continuous instruments and Z_i^d the remaining $q_z = (q_z^c + q_z^d)$ discrete instruments. For the r th component of Z_i^d , assume that Z_{ir}^d may take $(c_r - 1)$ distinct values, namely $Z_{ir}^d \in \{0, 1, \dots, c_r - 1\}$, for $r = 1, \dots, q_z^d$. Let $z = (z^c, z^d) \in \mathbb{R}^{q_z}$, $f(z) = f(z_r^c, z_r^d)$ denote the joint density function of (Z_i^c, Z_i^d) , and define $\mathcal{D} = \prod_{r=1}^{q_z^d} \{0, 1, \dots, c_r - 1\}$ as the support of Z_i^d .

For the continuous random variables, Z_i^c , we define the product kernel as

$$K_h(Z_i^c, z^c) = \prod_{s=1}^{q_z^c} \frac{1}{h_s} k\left(\frac{Z_{is}^c - z_s^c}{h_s}\right), \quad (3.4)$$

where $k(\cdot)$ is a univariate kernel function and $h = (h_1, \dots, h_{q_z^c})$ is the vector of bandwidths.

For the discrete regressors, we focus on the case in which no natural ordering exists, as in the binary case, and define

$$\ell(Z_{ir}^d, z_r^d, \lambda) = \begin{cases} 1 & \text{if } Z_{ir}^d = z_r^d \\ \lambda & \text{if } Z_{ir}^d \neq z_r^d, \end{cases} \quad (3.5)$$

where $\lambda \in [0, 1]$ is a smoothing parameter. Thus, $\ell(Z_i^d, z^d, 0)$ reduces to an indicator function when $\lambda = 0$, whereas $\ell(Z_i^d, z^d, \lambda)$ becomes a uniform weighting function when $\lambda = 1$. Define the indicator function $\mathbb{1}(Z_{ir}^d \neq z_r^d)$, which takes the value one when $(Z_{ir}^d \neq z_r^d)$, and zero otherwise. Moreover, let $d_{Z_i, z} = \sum_{r=1}^{q_z^d} \mathbb{1}(Z_{ir}^d \neq z_r^d)$ denote the number of mismatched components. The product kernel for the discrete

random variables is then given by

$$L(Z_i^d, z^d, \lambda) = \prod_{r=1}^{q_z^d} \ell(Z_{ir}^d, z_r^d, \lambda) = \mathbb{1}^{1-d_{Z_i, z}} \lambda^{d_{Z_i, z}}, \quad (3.6)$$

where $\lambda = (\lambda_1, \dots, \lambda_{q_z^d})$ is the vector of smoothing parameters.

Let $\phi = (h, \lambda) = (h_1, \dots, h_{q_z^c}, \lambda_1, \dots, \lambda_{q_z^d})$ collect the smoothing parameters associated with the continuous and discrete kernel components. The joint kernel weight function for mixed continuous-discrete variables is then defined as

$$\mathcal{W}_\phi(Z_i, z) = \prod_{s=1}^{q_z^c} \frac{1}{h} k\left(\frac{Z_{is}^c - z_s^c}{h}\right) \prod_{r=1}^{q_z^d} \ell(Z_{ir}^d, z_r^d, \lambda) = K_h(Z_i^c, z^c) L(Z_i^d, z^d, \lambda). \quad (3.7)$$

3.2.1 Estimation Procedure

First Stage

We begin by rewriting the partially linear model in a more compact form,

$$Y_i = m(X_{1i}) + V_i^\top \eta + \epsilon_i, \quad (3.8)$$

where $V_i = (X_{2i}, W_i^d)$, $m(\cdot)$ is an unknown smooth function, and ϵ_i is the structural error term. Following the partialling-out transformation of Robinson (1988), we take conditional expectations with respect to X_{1i} and subtract from both sides of (3.8), obtaining

$$Y_i - E[Y_i | X_{1i}] = \{V_i - E[V_i | X_{1i}]\}^\top \eta + \epsilon_i. \quad (3.9)$$

This transformation eliminates the nuisance function $m(X_{1i})$. Direct OLS estimation of (3.9) is nonetheless inconsistent, since $\text{Cov}(W_i^d, \epsilon_i) \neq 0$.

To deal with the endogeneity of W_i^d , we define the instrument vector $\Phi_i = (X_{2i}, p_i)$, where $p_i = E[W_i^d | Z_i] = P(W_i^d = 1 | Z_i)$. As discussed above, replacing W_i^d with p_i would amount to a forbidden regression and would therefore yield inconsistent second-stage estimates (Wooldridge 2010). By contrast, using p_i as an instrument

requires only the standard exclusion restriction $E[\epsilon_i | Z_i] = 0$, so that Φ_i is a valid instrument vector under $E[\epsilon_i | \Phi_i] = 0$.

The resulting infeasible IV estimator for η is given by

$$\tilde{\eta} = \left(\sum_{i=1}^N \tilde{\Phi}_i \tilde{V}_i^\top f^2(X_{1i}) \right)^{-1} \sum_{i=1}^N \tilde{\Phi}_i \tilde{Y}_i f^2(X_{1i}), \quad (3.10)$$

where $f(X_{1i})$ is the weighting density introduced to correct for the random denominator problem (Powell et al. 1989), and the transformed variables are defined by

$$\tilde{\Phi}_i = \Phi_i - E[\Phi_i | X_{1i}], \quad \tilde{V}_i = V_i - E[V_i | X_{1i}], \quad \tilde{Y}_i = Y_i - E[Y_i | X_{1i}],$$

with $E[\Phi_i | X_{1i}] = (E[X_{2i} | X_{1i}]^\top, E[p_i | X_{1i}]^\top)^\top$, and $E[V_i | X_{1i}] = (E[X_{2i} | X_{1i}]^\top, E[W_i^d | X_{1i}]^\top)^\top$.

The estimator $\tilde{\eta}$ is infeasible because it depends on unknown conditional expectations, the marginal density $f(X_{1i})$, and the unobserved probability p_i . These quantities, however, can all be estimated in order to construct a feasible estimator. In particular, p_i is replaced by $\hat{p}_i$, obtained from an estimator of $P(W_i^d = 1 | Z_i)$.

The literature on binary choice models offers a broad range of estimators, including parametric, semiparametric, and nonparametric alternatives. For estimation and simulation purposes, we consider four such estimators: probit, logit, the single-index estimator of Klein & Spady (1993), and the local likelihood logit of Frölich (2006). A detailed description of each method is provided in Appendix B.1. The conditional expectations $E[Y_i | X_{1i}]$, $E[X_{2i} | X_{1i}]$ and $E[\hat{p}_i | X_{1i}]$ are estimated using the following Nadaraya-Watson kernel regressions

$$\hat{E}[Y_i | X_{1i}] = \frac{\sum_{j=1}^N Y_j k\left(\frac{X_{1j} - X_{1i}}{h}\right)}{\sum_{j=1}^N k\left(\frac{X_{1j} - X_{1i}}{h}\right)}, \quad \hat{E}[X_{2i} | X_{1i}] = \frac{\sum_{j=1}^N X_{2j} k\left(\frac{X_{1j} - X_{1i}}{h}\right)}{\sum_{j=1}^N k\left(\frac{X_{1j} - X_{1i}}{h}\right)}, \quad \hat{E}[\hat{p}_i | X_{1i}] = \frac{\sum_{j=1}^N \hat{p}_j k\left(\frac{X_{1j} - X_{1i}}{h}\right)}{\sum_{j=1}^N k\left(\frac{X_{1j} - X_{1i}}{h}\right)}. \quad (3.11)$$

Meanwhile, $E[W_i^d | X_{1i}]$ is estimated by local-likelihood logit, and the marginal density $f(X_{1i})$ is estimated nonparametrically by a kernel density estimator, de-

noted as $\widehat{f}(X_{1i})$. The resulting feasible estimator is

$$\widehat{\eta} = \left(\sum_{i=1}^N \widehat{\Phi}_i \widehat{V}_i^\top \widehat{f}^2(X_{1i}) \right)^{-1} \sum_{i=1}^N \widehat{\Phi}_i \widehat{Y}_i \widehat{f}^2(X_{1i}), \quad (3.12)$$

where, for $\widehat{\Phi}_i = (X_{2i}, \widehat{p}_i)$, $\widehat{\Phi}_i = \widehat{\Phi}_i - \widehat{E}[\widehat{\Phi}_i \mid X_{1i}]$, $\widehat{V}_i = V_i - \widehat{E}[V_i \mid X_{1i}]$, and $\widehat{Y}_i = Y_i - \widehat{E}[Y_i \mid X_{1i}]$.

Second Stage

In the second stage, we turn to the estimation of the unknown function $m(\cdot)$, making use of the first-stage estimator $\widehat{\eta}$. Specifically, we rewrite the partially linear model in (3.8) as the following nonparametric model

$$Y_i^* = m(X_{1i}) + \epsilon_i^*, \quad (3.13)$$

where $Y_i^* = Y_i - V_i^\top \widehat{\eta}$ is the transformed dependent variable and $\epsilon_i^* = \epsilon_i - V_i^\top (\widehat{\eta} - \eta)$ is the corresponding transformed disturbance term.

Using a local linear nonparametric approach (see, among others, Fan & Gijbels (1992)), the resulting nonparametric estimator of $m(\cdot)$ is given by

$$\begin{aligned} \widehat{m}(x_1) &= e_1^\top \left[\sum_{i=1}^N K_h(X_{1i}, x_1) \begin{pmatrix} 1 & (X_{1i} - x_1)^\top \\ (X_{1i} - x_1) & (X_{1i} - x_1)(X_{1i} - x_1)^\top \end{pmatrix} \right]^{-1} \\ &\quad \times \sum_{i=1}^N K_h(X_{1i}, x_1) \begin{pmatrix} 1 \\ (X_{1i} - x_1) \end{pmatrix} Y_i^*, \end{aligned} \quad (3.14)$$

where $e_1 = (1, 0_{q_{x_1}})$ is the $(q_{x_1} + 1) \times 1$ vector whose first entry is equal to one and whose remaining entries are zero, and h denotes the bandwidth parameter. Derivations of the asymptotic properties are reported in Appendix B.2

3.2.2 Bandwidth Selection Procedure

Following Martins-Filho & Yao (2012), we adopt a Rule-of-Thumb (RoT) bandwidth selection procedure throughout the estimation stages. Although this choice simplifies implementation, it should be noted that selecting an optimal bandwidth in the first stage does not necessarily guarantee optimal performance in subsequent stages. Nonetheless, the use of a common bandwidth selection rule facilitates comparability across stages (Su & Ullah 2008; Delgado & Parmeter 2014).

More precisely, the optimal bandwidth for each regression type is assumed to be proportional to sample size and is given by $h \sim N^{-\frac{1}{4+q_c}}$ and $\lambda \sim N^{-\frac{2}{4+q_c}}$, where q_c denotes the number of continuous regressors under consideration.

For continuous covariates, the bandwidth is further rescaled by the variability of the covariate, yielding $h = c \sigma_X N^{-\frac{1}{4+q_c}}$ where c is a scale factor and σ_X denotes the standard deviation of the covariate. For discrete covariates, in the absence of a formal rule, we follow the practical recommendation in Henderson & Parmeter (2015) and adopt a pilot bandwidth based on sample size $\lambda = c N^{-\frac{2}{4+q_c}}$.

3.3 Monte Carlo Simulations

This section presents two complementary Monte Carlo experiments designed to evaluate the finite-sample performance of the proposed IV estimators in the presence of endogenous discrete regressors. Although both experiments are based on the same data-generating processes (DGPs), they are intended to address distinct questions and therefore rely on different evaluation criteria.

The first experiment follows a conventional simulation framework and focuses on standard finite-sample properties. In particular, it examines whether the estimator behaves in finite samples as predicted by the asymptotic theory, using bias, standard deviation (Sd), and RMSE as the primary performance measures.

The second experiment adopts an out-of-sample perspective, following Gao et al. (2015). Rather than emphasizing repeated-sample accuracy, it evaluates predictive performance under possible model misspecification. More specifically, estimators are trained on one sample and then assessed on an independent hold-out sample using mean squared prediction error, thereby allowing for a comparison across alternative first- and second-stage specifications.

3.3.1 Finite In-Sample Performance

To investigate the finite-sample properties of the proposed estimator, we conduct a Monte Carlo simulation based on the following DGPs,

$$\text{DGP1} : Y_i = -\frac{1}{2}X_{1i} + \sin(2X_{1i}) + 4\left(\frac{1}{1 + \exp(-3(X_{1i} - 1))}\right) + \beta X_{2i} + \gamma W_i^d + \epsilon_i, \quad (3.15)$$

$$\text{DGP2} : Y_i = \frac{1}{2}X_{1i} + 4\sin(2X_{1i}) - X_{1i}^2 + \beta X_{2i} + \gamma W_i^d + \epsilon_i. \quad (3.16)$$

The explanatory variables are generated as $(X_{1i}, X_{2i}) \sim^{i.i.d.} \mathcal{U}(-5, 5)$, and the slope parameters are fixed at $\beta = \gamma = 1$. The instruments $Z_i = (Z_{1i}, Z_{2i})$ are independently generated according to $Z_i \sim^{i.i.d.} \mathcal{U}(-5, 5)$. To induce endogeneity in W_i^d , we specify the latent index as

$$W_i^{d*} = c_0 + c_1 Z_{1i} + Z_{1i}^2 + \sin(Z_{2i}) + u_i, \quad (3.17)$$

$$W_i^d = \mathbb{1}(W_i^{d*} > 0), \quad (3.18)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. The constant c_0 acts as a threshold-shifting term in the latent index as is calibrated so that $\Pr(W_i^d = 1) \approx \Pr(W_i^d = 0) \approx 0.5$. The coefficient c_1 determines instrument strength: larger values imply a stronger first stage, with $c_1 = 1$ taken as the baseline case, whereas values satisfying $0 < c_1 < 1$ correspond to weaker instruments.

We further assume that ϵ_i and u_i are jointly normally distributed according to

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \sim i.i.d. \quad N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\right) \quad (3.19)$$

where $\theta = (0.2, 0.5, 0.8)$ governs the degree of endogeneity. The number of Monte Carlo replications is set to 1,000. We employ the Epanechnikov kernel function $k(u) = 0.75(1 - u^2)\mathbb{1}\{|u| \leq 1\}$ and consider sample sizes $N \in \{100, 200, 400\}$.

To evaluate the performance of the estimator proposed for η in (3.12), we report bias, Standard deviation (Sd), and Root Mean Squared Error (RMSE). For

the nonparametric estimator of $m(\cdot)$ in (3.14), performance is assessed using the RMSE. To ensure robust and reliable results, the analysis is restricted to the 10%-90% quantile range of the simulation outcomes, thereby excluding extreme values (Martins-Filho & Yao 2012; Delgado & Parmeter 2014).

Tables 3.1, 3.2, 3.3, and 3.4 report the Monte Carlo simulation results for the discrete case, evaluating the finite-sample performance of the proposed estimator of $\widehat{\gamma}$ under both DGP1 and DGP2.

Figures 3.1, 3.2, 3.3, and 3.4 present the RMSE distributions of the nonparametric estimator $\widehat{m}(\cdot)$.

The results reveal substantial differences across first-stage estimators. Under probit and logit first-stage specifications, $\widehat{\gamma}$ exhibits considerable dispersion and noticeable instability in small samples, with performance deteriorating further under DGP2. Although precision improves markedly as N increases, these findings suggest that parametric first-stage estimators require relatively larger samples to deliver reliable estimates. By contrast, the single index and local likelihood logit estimators generate stable and well-behaved estimates across all sample sizes and under both DGPs, with negligible bias and steadily declining dispersion as N increases.

A similar pattern emerges for the nonparametric estimator $\widehat{m}(\cdot)$. Under probit and logit first stages, the RMSE distributions display substantial dispersion and pronounced upper tails at small sample sizes, especially under DGP2. In contrast, the Single Index and local logit first-stage estimators yield tightly concentrated RMSE distributions throughout. In all cases, both bias and variance decline as the sample size increases, in accordance with the theoretical properties established for the parametric and nonparametric components of the estimator.

Table 3.1: Finite sample performance of $\widehat{\gamma}$ with probit $\widehat{p}$

θ	N	DGP1			DGP2		
		<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>	<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>
0.2	100	-0.2069	6.4142	6.4175	-0.9886	14.3960	14.4299
	200	-0.0944	6.1066	6.1073	0.0232	8.7532	8.7533
	400	-0.0360	0.7690	0.7698	-0.0975	1.8565	1.8591
0.5	100	-0.2153	6.2134	6.2171	-1.0537	16.3069	16.3409
	200	-0.1388	7.5566	7.5579	-0.0545	10.7087	10.7089
	400	-0.0340	0.7673	0.7681	-0.0912	1.8400	1.8422
0.8	100	-0.1659	7.5443	7.5462	-0.8893	22.9633	22.9806
	200	-0.1776	7.9336	7.9356	-0.1207	11.0795	11.0801
	400	-0.0314	0.7778	0.7785	-0.0918	1.8566	1.8589

Note: The table reports average Bias, standard deviation (Sd), and RMSE of $\widehat{\gamma}$ over 1,000 replications for the three-stage estimator under DGP1 and DGP2.

Table 3.2: Finite sample performance of $\widehat{\gamma}$ with logit $\widehat{p}$

θ	N	DGP1			DGP2		
		<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>	<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>
0.2	100	-0.2067	6.4129	6.4162	-0.9799	14.2676	14.3012
	200	-0.0937	6.0797	6.0804	0.0243	8.7171	8.7171
	400	-0.0358	0.7661	0.7669	-0.0970	1.8498	1.8523
0.5	100	-0.2156	6.2143	6.2181	-1.0471	16.2734	16.3070
	200	-0.1414	7.6289	7.6302	-0.0578	10.7982	10.7984
	400	-0.0339	0.7644	0.7651	-0.0907	1.8332	1.8354
0.8	100	-0.1674	7.4906	7.4925	-0.889	22.7629	22.7802
	200	-0.1803	8.0089	8.0109	-0.1243	11.1721	11.1727
	400	-0.0313	0.7749	0.7756	-0.0914	1.8499	1.8521

Note: The table reports average Bias, standard deviation (Sd), and RMSE of $\widehat{\gamma}$ over 1,000 replications for the three-stage estimator under DGP1 and DGP2.

Table 3.3: Finite sample performance of $\widehat{\gamma}$ with single index $\widehat{p}$

θ	N	DGP1			DGP2		
		<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>	<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>
0.2	100	0.0473	0.2933	0.2971	0.0327	0.7956	0.7963
	200	0.0291	0.1888	0.1911	0.0334	0.5004	0.5015
	400	0.0077	0.1304	0.1307	0.0055	0.2933	0.2933
0.5	100	0.0717	0.2943	0.3029	0.0581	0.7971	0.7992
	200	0.0414	0.1885	0.193	0.0444	0.4985	0.5005
	400	0.0160	0.1311	0.1321	0.0147	0.2937	0.2941
0.8	100	0.0925	0.2922	0.3065	0.0733	0.7939	0.7973
	200	0.0531	0.1886	0.1959	0.0592	0.4974	0.5009
	400	0.0257	0.1322	0.1346	0.0253	0.2953	0.2964

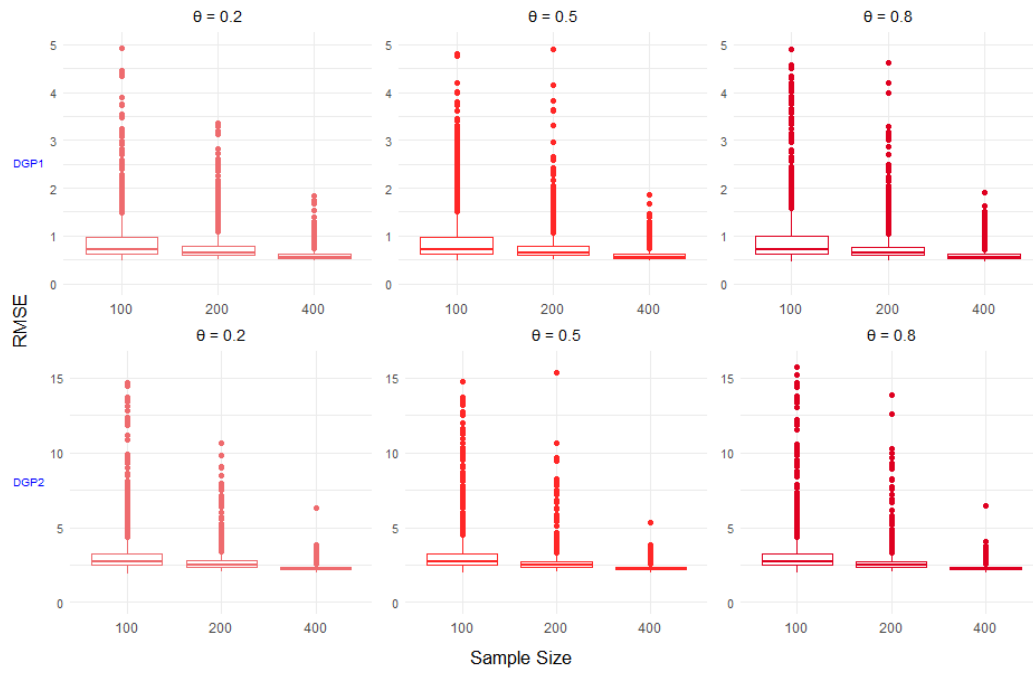
Note: The table reports average Bias, standard deviation (Sd), and RMSE of $\widehat{\gamma}$ over 1,000 replications for the three-stage estimator under DGP1 and DGP2.

Table 3.4: Finite sample performance of $\widehat{\gamma}$ with local likelihood logit $\widehat{p}$

θ	N	DGP1			DGP2		
		<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>	<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>
0.2	100	0.0375	0.2932	0.2956	0.0123	0.8153	0.8154
	200	0.0271	0.1867	0.1887	0.0306	0.4944	0.4953
	400	0.0161	0.1271	0.1281	0.0146	0.2828	0.2832
0.5	100	0.0504	0.2899	0.2943	0.0233	0.8111	0.8114
	200	0.0375	0.1861	0.1899	0.0415	0.4937	0.4955
	400	0.0220	0.1270	0.1289	0.0212	0.2839	0.2847
0.8	100	0.0667	0.2862	0.2939	0.0416	0.8091	0.8102
	200	0.0468	0.1864	0.1922	0.0497	0.4931	0.4956
	400	0.0274	0.1272	0.1301	0.0269	0.2845	0.2858

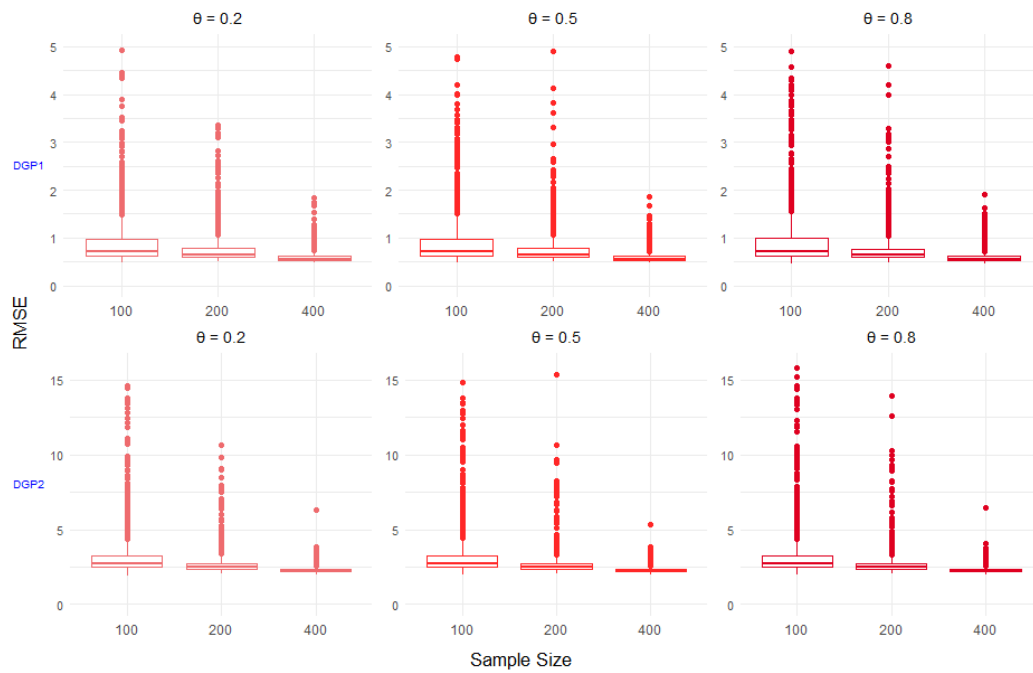
Note: The table reports average Bias, standard deviation (Sd), and RMSE of $\widehat{\gamma}$ over 1,000 replications for the three-stage estimator under DGP1 and DGP2.

Figure 3.1: Finite sample performance of $\widehat{m}(\cdot)$: probit first stage



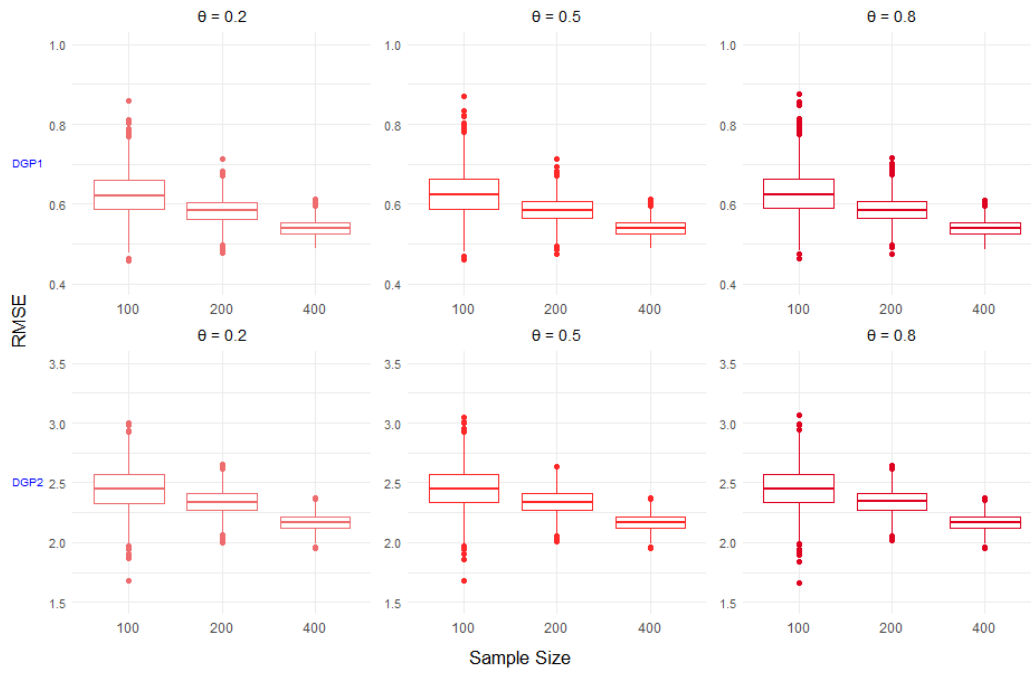
Note: Boxplots of the RMSE values of the nonparametric estimator of $m(\cdot)$ in 1,000 replications

Figure 3.2: Finite sample performance of $\widehat{m}(\cdot)$: logit first stage



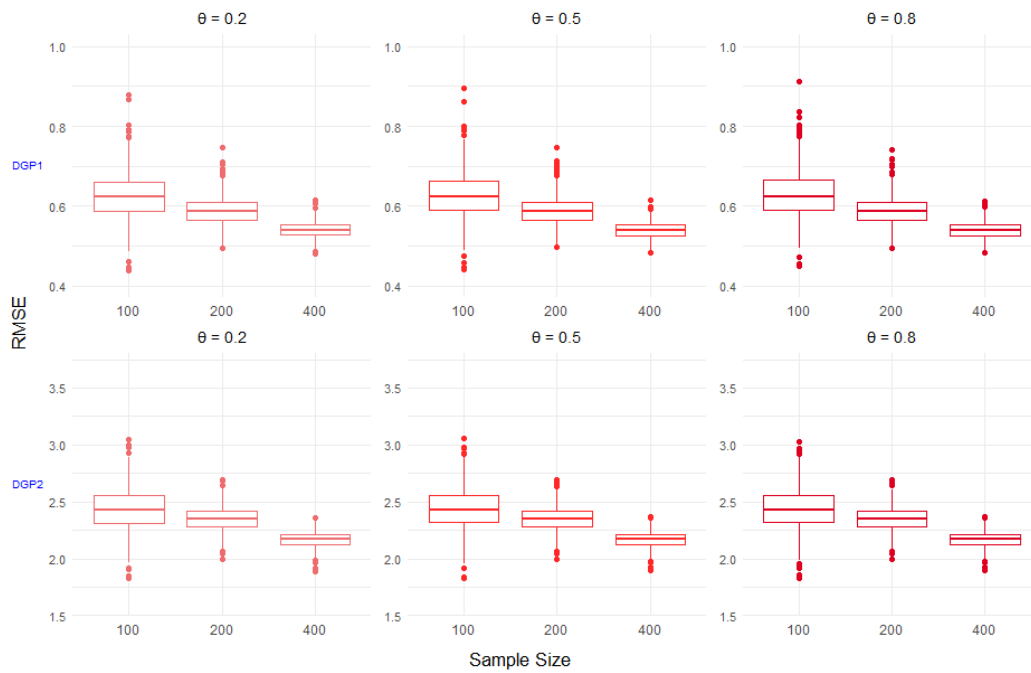
Note: Boxplots of the RMSE values of the nonparametric estimator of $m(\cdot)$ in 1,000 replications

Figure 3.3: Finite sample performance of $\widehat{m}(\cdot)$: single index first stage



Note: Boxplots of the RMSE values of the nonparametric estimator of $m(\cdot)$ in 1,000 replications

Figure 3.4: Finite sample performance of $\widehat{m}(\cdot)$: local likelihood logit first stage



Note: Boxplots of the RMSE values of the nonparametric estimator of $m(\cdot)$ in 1,000 replications

3.3.2 Finite Out-of-Sample Performance

In this section, we extend the previous Monte Carlo experiment to evaluate the finite-sample performance of the proposed IV estimators from an out-of-sample predictive perspective. To compare the estimators rigorously, and following Gao et al. (2015), we assess their performance on independent hold-out samples. Specifically, we compute the out-of-sample mean squared error as $\text{MSE} = n_2^{-1} \sum_{i=1}^{n_2} (Y_i - \widehat{Y}_i)^2$, where Y_i and $\widehat{Y}_i$ denote the observed and predicted values, respectively, for an independent evaluation sample of size n_2 .

The true GDP is given by

$$Y_i = \beta_1 X_{1i} + \beta_2 (\sin(X_{1i})) + \beta_3 \left(\frac{1}{1 + \exp(-3(X_{1i} - 1))} \right) + \beta_4 X_{2i} + \gamma W_i^d + \epsilon_i, \quad (3.20)$$

where $W_i^d = \mathbb{1}\{c_0 + c_1 Z_{1i} + \sin(X_{1i}) + X_{2i} + X_{2i}^2 + u_i > 0\}$. Against this benchmark, we compare the following estimators:

1. The structural equation in (3.20) is correctly specified, and $P(W_i^d = 1 \mid Z_{1i}, X_{1i}, X_{2i})$ is estimated using standard probit (Oracle-Probit) and logit (Oracle-Logit) estimators.
2. The structural equation is misspecified as a fully parametric (PAR) model,

$$Y_i = \alpha X_{1i} + \beta X_{2i} + \gamma W_i^d + u_i \quad (3.21)$$

where $W_i^d = \mathbb{1}\{\pi_0 Z_{1i} + \pi_1 X_{1i} + \pi_2 X_{2i} > 0\}$. Within this class, $P(W_i^d = 1 \mid Z_{1i}, X_{1i}, X_{2i})$ is estimated using

- a) parametric probit (Probit-PAR) and logit (Logit-PAR) estimators;
 - b) semiparametric single index estimator (SP-PAR) of Klein & Spady (1993);
 - c) nonparametric (NP) local likelihood logit (NP-PAR) of Frölich (2006).
3. The structural equation is specified as a semiparametric partially linear

model

$$Y_i = m(X_1) + \beta X_{2i} + \gamma W_i^d + u_i, \quad (3.22)$$

where $W_i^d = \mathbb{1}\{\pi_0 Z_{1i} + \pi_1 X_{1i} + \pi_2 X_{2i} > 0\}$. Within this class, $P(W_i^d = 1 \mid Z_{1i}, X_{1i}, X_{2i})$ is estimated using

- a) parametric probit (Probit-PLM) and logit (Logit-PLM) estimators.
- b) semiparametric single index estimator (SP-PLM) of Klein & Spady (1993);
- c) nonparametric (NP) local likelihood logit (NP-PLM) estimator of Frölich (2006).

We generate 1,000 Monte Carlo replications from the DGP. In each replication, the models are estimated on a training sample of size n_1 and evaluated on an independent hold-out sample of size $n_2 = 1,000$ using the mean squared prediction error. To examine how predictive performance varies with sample size, we consider training samples of size $n_1 \in \{100, 200, 400\}$. The endogeneity parameter is fixed at $\theta = 0.65$, which lies between the intermediate and high levels of endogeneity considered in the previous section.

The results, reported in terms of the mean, median, Standard deviation (Sd), and interquantile range (IQR) of the Mean Squared Error (MSE), are summarized in Table 3.5.

Table 3.5: Monte Carlo results

n_1	Model	Median (MSE)	Mean (MSE)	SD (MSE)	IQR (MSE)
100	Oracle–Logit	1.152	1.921	3.262	0.404
	Oracle–Probit	1.152	1.923	3.262	0.411
	NP–PLM	1.681	1.681	0.101	0.134
	SP–PLM	1.683	1.682	0.103	0.132
	Probit–PLM	1.900	3.250	5.646	0.901
	Logit–PLM	1.899	3.246	5.646	0.881
	NP–PAR	1.946	1.952	0.099	0.118
	SP–PAR	1.949	1.954	0.100	0.119
	Logit–PAR	2.243	3.716	9.335	0.841
	Probit–PAR	2.245	3.720	9.339	0.847
200	Oracle–Probit	1.070	1.334	1.445	0.139
	Oracle–Logit	1.072	1.333	1.443	0.137
	NP–PLM	1.582	1.584	0.083	0.109
	SP–PLM	1.586	1.586	0.083	0.108
	Logit–PLM	1.728	2.071	1.240	0.358
	Probit–PLM	1.729	2.072	1.239	0.360
	NP–PAR	1.915	1.920	0.085	0.113
	SP–PAR	1.918	1.922	0.084	0.115
	Logit–PAR	2.036	2.373	0.988	0.301
	Probit–PAR	2.039	2.375	0.990	0.298
400	Oracle–Logit	1.055	1.120	0.282	0.117
	Oracle–Probit	1.055	1.121	0.283	0.117
	NP–PLM	1.477	1.479	0.065	0.084
	SP–PLM	1.479	1.480	0.066	0.085
	Logit–PLM	1.558	1.688	0.510	0.159
	Probit–PLM	1.559	1.689	0.510	0.160
	NP–PAR	1.907	1.902	0.085	0.107
	SP–PAR	1.912	1.903	0.085	0.107
	Logit–PAR	1.978	2.118	0.481	0.184
	Probit–PAR	1.978	2.119	0.482	0.185

As expected, the oracle estimators (Oracle–Logit and Oracle–Probit), which exploit knowledge of the true specification in both stages, achieve the lowest out-of-sample MSE across all sample sizes; these estimators are, of course, infeasible in practice. Among feasible estimators, the fully parametric specifications (Logit–PAR and Probit–PAR) are uniformly dominated at every sample size, confirming

their vulnerability to misspecification. The partially linear specifications outperform their parametric counterparts, indicating that the semiparametric structure yields substantial bias reduction at only a modest variance cost.

Within the partially linear class, the semiparametric and nonparametric first-stage variants (SP-PLM and NP-PLM) stand out: they achieve the lowest feasible MSE and exhibit markedly lower dispersion—standard deviations and interquartile ranges roughly one-fifth to one-half those of Logit-PLM and Probit-PLM—while converging more rapidly toward the oracle benchmark as n_1 grows. The parametric first-stage variants, by contrast, display heavy right tails, particularly at $n_1 = 100$, where the mean MSE substantially exceeds the median (e.g., 3.250 vs. 1.900 for Probit-PLM), signalling occasional large prediction errors.

3.4 Revisiting the Knowledge Capital Production Function

Technical change (TC) has long been recognized as a fundamental driver of economic development, making the assessment of its impact on the production process a central concern in both theoretical and empirical research. At the firm level, a prominent strand of the literature builds on the knowledge capital production function framework introduced by Pakes & Griliches (1984) & Griliches (1998), which posits a causal relationship between unobservable knowledge capital and observable indicators of innovation activity, such as R&D expenditures, patents, and firm-level performances. Building on this framework, Crépon et al. (1998) proposed the so-called CDM model that is a simultaneous three-equation system closely aligned with Griliches' original formulation and that has become highly influential and extensively extended in subsequent research (Löf et al. 2017).

The increasing availability of harmonized Community Innovation Surveys (CIS) has facilitated the construction of direct binary measures of innovation output, which have emerged as the most widely used alternative to continuous proxies (Griffith et al. 2006; Parisi et al. 2006; Duguet 2006; Masso & Vahter 2008; Chudnovsky et al. 2006; Musolesi & Huiban 2010; Mairesse & Mohnen 2010; Mohnen & B. H. Hall 2013). These measures enhance comparability across studies and survey waves, thereby improving the consistency of empirical analyses (B. H. Hall, Lotti, et al. 2009; B. H. Hall 2011).

With respect to the choice of innovation variable, this study focuses on process innovation (IP), rather than alternative indicators such as product innovation or broader innovation measures. A growing body of evidence (Jaumandreu & Mairesse 2010; Jaumandreu & Mairesse 2017; Jaumandreu & Lin 2018) suggests that process innovation primarily enhances productivity through reductions in production costs, whereas product innovation operates mainly through shifts in demand. This distinction is further supported by the literature on technology and employment, which documents asymmetric labour-market effects of the two types of innovation (Vivarelli 2014; Harrison et al. 2014; Calvino & Virgillito 2018). In particular, concerns regarding technological unemployment are typically associated with process innovation, while product innovation tends to exert labor-friendly effects through the expansion of markets and demand (Bogliacino et al. 2012).

Regarding functional form, the standard empirical approach relies on the Cobb-Douglas production function, which imposes strong *a priori* restrictions, including unitary elasticities of substitution and separability across inputs (Cobb & Douglas 1928). A key implication of this specification is that the additive inclusion of a binary innovation variable implies a neutral effect on the production process, equivalent to multiplicative separability between the innovation com-

ponent and the input-based production function (Uzawa 1961), in line with the neoclassical classification of technical change proposed by Hicks (1932); see Antonioli et al. (2021) for a detailed discussion in the context of production function estimation. Under these assumptions, we specify the following Cobb-Douglas production function,

$$y_i = \beta_0 + \beta_k k_i + \beta_\ell \ell_i + \gamma IP_i + \varepsilon_i, \quad (3.23)$$

where y_i denotes log value added, k_i and ℓ_i represent log capital and log labor, respectively, IP_i is a binary indicator of process innovation, and sectoral fixed effects are included throughout.

With respect to endogeneity, the innovation variable is treated as endogenous, and innovation inputs are employed as instruments. For expositional purposes, standard inputs are assumed exogenous, consistent with empirical evidence indicating that OLS provides reliable estimates relative to IV-based alternatives in value-added production functions (Musolesi & Huiban 2010; Antonioli et al. 2021), as well as with the findings of D. Akerberg et al. (2007), who show that transmission bias is not empirically relevant in this context and that OLS performs comparably, or even better, than standard fixed effects estimators.

Given the restrictive nature of the Cobb-Douglas specification, subsequent contributions have sought to relax these assumptions by adopting more flexible functional forms (Christensen et al. 1973; Gallant 1981; Ivaldi et al. 1996). A further strand of the literature has moved beyond parametric restrictions altogether, employing nonparametric and semiparametric approaches (Henderson & Russell 2005; Henderson & Kumbhakar 2006; Kumbhakar, Park, et al. 2007; Kumbhakar, Sun, et al. 2016; Ma et al. 2020). Following this strand of the literature, we specify the following partially linear model,

$$y_i = m(k_i, \ell_i) + \gamma IP_i + \epsilon_i, \quad (3.24)$$

where $m(\cdot)$ is an unknown smooth function of the continuous inputs, estimated nonparametrically, and all other terms are defined as in equation (3.23).

The empirical analysis is based on firm-level data drawn from the tenth Unicredit–Capitalia survey, merged with balance sheets from AIDA or the Chambers of Commerce (Antonioli et al. 2021). Value added, capital, and labor are measured in natural logarithms for the year 2006. The process innovation variable takes the value one if a firm reports at least one process innovation during the period 2004–2006. Manufacturing firms are classified into 22 sectors according to the two-digit ATECO2002 classification.

Instrumental variables are selected from the determinants of innovation, following Musolesi & Huiban (2010) and the multi-step procedure proposed by Antonioli et al. (2021), which combines measures of knowledge inputs and absorptive capacity. The final set of instruments includes *BetterProd* (investment in product improvement), *FinanIncent* (receipt of fiscal or financial incentives), and *MarkPenEU15* (market penetration in EU15). After applying these selection criteria, the baseline sample consists of 2924 firms. When restricting the analysis to IV specifications, the sample size is reduced to 2,598 observations due to missing values in the instrument set (see Appendix B.3 for further details).

Table 3.6 reports the estimation results for the parametric specification in (3.23). We compare OLS estimates with IV approaches based on alternative first-stage estimators used to construct fitted probabilities, including probit (IV-P), logit (IV-L), the single-index estimator (IV-SP), and the local likelihood logit estimator (IV-NP).

Table 3.6: Parametric structural model estimates

	OLS (i)	P-IV (ii)	L-IV (iii)	SP-IV (iv)	NP-IV (v)
β_ℓ	0.7500 (0.0137)	0.7295 (0.0198)	0.7294 (0.0198)	0.7335 (0.0194)	0.7458 (0.0180)
β_k	0.1849 (0.0092)	0.1791 (0.0123)	0.1791 (0.0124)	0.1809 (0.0120)	0.1866 (0.0118)
γ	0.0196 (0.0151)	0.2828 (0.1031)	0.2824 (0.1032)	0.2997 (0.0978)	0.2635 (0.0903)
E.S.	0.9349	0.9086	0.9085	0.9144	0.9324
MRTS	4.0562	4.0728	4.0731	4.0538	3.9960
N	2924	2598	2598	2598	2598

Notes: Standard errors are reported in parentheses. E.S. denotes elasticity of scale ($\hat{\beta}_\ell + \hat{\beta}_k$), while MRTS denotes the marginal rate of technical substitution ($\hat{\beta}_\ell/\hat{\beta}_k$). The IP variable is treated as endogenous in columns (ii)–(v), where alternative generated instruments are employed; in contrast, OLS treats IP as exogenous. P-IV and L-IV rely on fitted probabilities $\hat{p}$ obtained from probit and logit first-stage estimations, respectively; SP-IV is based on the Klein–Spady single-index estimator; and NP-IV employs the nonparametric first-stage procedure of Frölich (2006), using a common pilot bandwidth $h = 1.06 \times n^{-1/4} = 0.1485$ applied to both the instrument and the sectoral variable.

While the OLS estimate of γ is small and statistically insignificant ($\hat{\gamma} = 0.020$), all IV specifications yield positive and statistically significant estimates, with $\hat{\gamma}$ ranging between approximately 0.26 and 0.30. The estimated coefficients on labor and capital remain remarkably stable across IV approaches, with $\hat{\beta}_\ell \approx 0.73$ and $\hat{\beta}_k \approx 0.18$, implying an elasticity of scale in the range of 0.91–0.93.

We next consider the semiparametric structural model in (3.24). Table 3.7 reports the corresponding estimates using the same set of first-stage alternatives. The PLM-OLS estimate of γ is again small and not statistically significant ($\hat{\gamma} = 0.007$), whereas all IV specifications produce positive and significant estimates. When fitted probabilities are obtained from parametric and semiparametric first stages, $\hat{\gamma}$ lies in the range 0.40–0.42, while the PLM-NP-IV specification yields a lower estimate, $\hat{\gamma} = 0.337$, consistent with the systematically lower NP-IV estimates observed in the parametric results.

The estimated input elasticities remain broadly stable across specifications. In particular, $\hat{\beta}_\ell$ is approximately 0.62 and $\hat{\beta}_k$ is approximately 0.19, implying an elasticity of scale close to 0.81. Moving from OLS to IV estimation results in a

modest downward adjustment of both coefficients: $\hat{\beta}_\ell$ decreases from 0.635 to approximately 0.616–0.621, while $\hat{\beta}_k$ declines from 0.199 to about 0.191–0.193, reducing the average elasticity of scale from 0.833 to roughly 0.807–0.814. Notably, the PLM-NP-IV specification produces a substantially smaller standard error for $\hat{\gamma}$ (0.115 compared with approximately 0.174), reflecting the smoother fitted probabilities generated by the nonparametric first stage.

Table 3.8 presents a sensitivity analysis with respect to the sectoral bandwidth in the NP first stage, holding the bandwidth for IVs fixed. The estimate of $\hat{\gamma}$ increases monotonically from 0.317 at $h = 0.1$ to 0.420 at $h = 0.9$. When the bandwidth is small, fitted probabilities are estimated separately by sector, effectively absorbing sectoral heterogeneity in the first stage and resulting in a lower estimate of γ . As the bandwidth increases, sectors are progressively pooled, and $\hat{\gamma}$ rises to capture variation that is no longer accounted for in the first stage.

Finally, Tables 3.9 and 3.10 report the distribution of pointwise input elasticity estimates and elasticities of scale. Under PLM-OLS, the nonparametric component generates substantial firm-level heterogeneity, with mean capital and labour elasticities of 0.199 and 0.635, respectively, and an average elasticity of scale of 0.833. Across IV specifications, this heterogeneity persists, with mean elasticities of capital and labour elasticities in the ranges 0.191–0.193 and 0.616–0.620, respectively, accompanied by considerable dispersion around these averages.

Table 3.7: Nonparametric structural model estimates

	PLM-OLS (i)	PLM-P-IV (ii)	PLM-L-IV (iii)	PLM-SP-IV (iv)	PLM-NP-IV (v)
β_ℓ	0.6346 (0.0137)	0.6161 (0.0145)	0.6161 (0.0144)	0.6167 (0.0144)	0.6212 (0.0140)
β_k	0.1986 (0.0092)	0.1911 (0.0084)	0.1911 (0.0084)	0.1913 (0.0083)	0.1931 (0.0078)
γ	0.0072 (0.0186)	0.4166 (0.1735)	0.4165 (0.1737)	0.4036 (0.1591)	0.3373 (0.1145)
E.S.	0.8332	0.8072	0.8072	0.8080	0.8143
MRTS	4.0562	3.2241	3.2241	3.2233	3.2176
N	2924	2598	2598	2598	2598

Notes: Standard errors are reported in parentheses. E.S. denotes the elasticity of scale evaluated at the mean. PLM-OLS treats the innovation variable as exogenous. PLM-P-IV and PLM-L-IV rely on fitted probabilities $\hat{p}$ obtained from probit and logit first-stage estimations, respectively, while PLM-SP-IV is based on the Klein-Spady single-index estimator. For columns (i)–(iv), the bandwidth associated with the sectoral discrete variable is set to 0.1. PLM-NP-IV employs the nonparametric first-stage procedure of Frölich (2006), using a common pilot bandwidth $h = 1.06 \times n^{-1/4} = 0.1485$ applied to both the instrument and the sectoral variable. In all cases, the first-stage estimation excludes capital and labor to mitigate potential endogeneity arising from input variables.

Table 3.8: PLM-NP-IV estimates: sensitivity to Sector Bandwidth

	$h = 0.1$	$h = 0.2$	$h = 0.3$	$h = 0.4$	$h = 0.5$	$h = 0.6$	$h = 0.7$	$h = 0.8$	$h = 0.9$
$\hat{\gamma}$	0.3169 (0.1060)	0.3530 (0.1216)	0.3741 (0.1315)	0.3880 (0.1385)	0.3982 (0.1437)	0.4058 (0.1477)	0.4117 (0.1509)	0.4165 (0.1535)	0.4204 (0.1557)
$\hat{\beta}_k$	0.1929 (0.0079)	0.1922 (0.0081)	0.1919 (0.0082)	0.1916 (0.0082)	0.1914 (0.0083)	0.1913 (0.0083)	0.1912 (0.0084)	0.1911 (0.0084)	0.1910 (0.0084)
$\hat{\beta}_\ell$	0.6206 (0.0141)	0.6190 (0.0142)	0.6180 (0.0143)	0.6174 (0.0143)	0.6169 (0.0144)	0.6166 (0.0144)	0.6163 (0.0144)	0.6161 (0.0145)	0.6159 (0.0145)
E.S.	0.8135	0.8112	0.8099	0.8090	0.8083	0.8078	0.8075	0.8072	0.8069
N	2598								

Notes: Standard errors are reported in parentheses. Each column presents PLM-IV-NP estimates based on the nonparametric first-stage procedure of Frölich (2006), where the bandwidth h applied to the sectoral discrete variable is varied, while the instrument bandwidth is held fixed at $h_{iv} = 0.1485$. The input elasticities $\hat{\beta}_k$ and $\hat{\beta}_\ell$ are computed as sample averages of the corresponding pointwise estimates. E.S. denotes mean elasticity of scale.

Table 3.9 presents the distribution of firm-level (pointwise) input elasticity estimates. Across all IV specifications—PLM-IV-Probit/Logit, PLM-IV-SP, and PLM-IV-NP—the nonparametric component $m(k, \ell)$ captures substantial heterogeneity. In particular, the interquartile range of $\hat{\beta}_k$ lies approximately within $[0.125, 0.248]$, while that of $\hat{\beta}_\ell$ spans roughly $[0.518, 0.751]$. Table 3.10 reports the corresponding elasticities of scale. When evaluated at the sample mean of the inputs, $(\bar{k}, \bar{\ell})$, the estimated elasticity of scale $\widehat{\text{ES}}(\bar{k}, \bar{\ell}) \approx 0.917\text{--}0.939$, indicates mild decreasing returns to scale. By contrast, the mean of the pointwise estimates, $\overline{\text{ES}} \approx 0.807\text{--}0.833$, is lower, reflecting the influence of firms operating in regions of the input

space characterized by weaker returns to scale.

A key advantage of the partially linear specification is its ability to accommodate firm-level heterogeneity in input elasticities. Rather than imposing a constant effect, the model yields a full distribution of marginal effects across firms. The distribution of capital elasticities (Figure 3.5), centered around a median of 0.195 and a mean of 0.199, is right-skewed, with indications of multimodality near 0.10 and a long upper tail extending beyond 0.60. This pattern points to substantial heterogeneity in the marginal productivity of capital. In contrast, the distribution of labor elasticities (Figure 3.6) is unimodal, with a median of 0.666 and a mean of 0.635. Most of the mass is concentrated between 0.4 and 0.9, and the distribution exhibits mild left skewness, with a longer tail toward lower values.

Table 3.9: Estimates of input elasticity: mean and quantiles

Method	Capital					Labour				
	$\bar{\beta}_k$	$\hat{\beta}_k(\bar{x})$	$\hat{\beta}_{k,25}$	$\hat{\beta}_{k,50}$	$\hat{\beta}_{k,75}$	$\bar{\beta}_\ell$	$\hat{\beta}_\ell(\bar{x})$	$\hat{\beta}_{\ell,25}$	$\hat{\beta}_{\ell,50}$	$\hat{\beta}_{\ell,75}$
PLM-OLS	0.1986	0.2427	0.1361	0.1948	0.2534	0.6346	0.6746	0.5328	0.6664	0.7670
PLM-IV-Probit/Logit	0.1911	0.2307	0.1249	0.1854	0.2481	0.6161	0.7082	0.5177	0.6500	0.7512
PLM-IV-SP	0.1913	0.2311	0.1249	0.1865	0.2475	0.6167	0.7071	0.5195	0.6513	0.7515
PLM-IV-NP	0.1925	0.2330	0.1264	0.1923	0.2468	0.6197	0.7017	0.5187	0.6546	0.7498

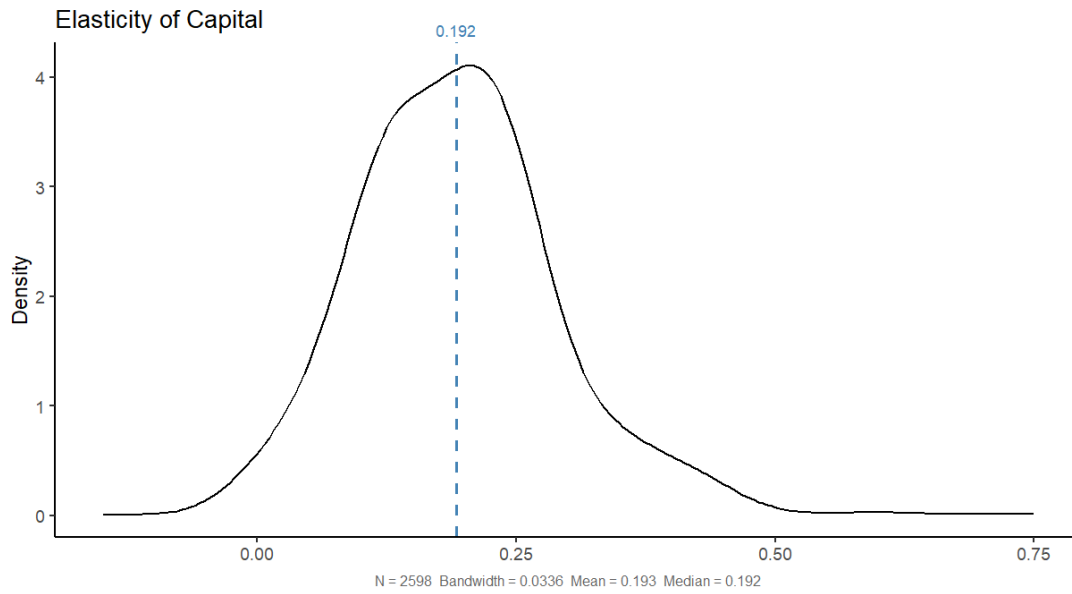
Notes: $\bar{\beta}$ denotes the sample mean of the pointwise elasticity estimates, while $\hat{\beta}(\bar{x})$ denotes the elasticity evaluated at the sample mean of the input variables. Reported quantiles refer to the empirical distribution of the pointwise elasticity estimates across firms. Probit and logit specifications yield identical estimates and are therefore reported jointly in a single row.

Table 3.10: Elasticities of scale: mean and quantiles

Method	$\overline{ES}$	$\widehat{ES}(\bar{x})$	ES_{25}	ES_{50}	ES_{75}
PLM-OLS	0.8332	0.9173	0.7212	0.8583	0.9626
PLM-IV-Probit/Logit	0.8072	0.9389	0.7024	0.8176	0.9353
PLM-IV-SP	0.8080	0.9382	0.7042	0.8187	0.9360
PLM-IV-NP	0.8122	0.9347	0.7101	0.8257	0.9362

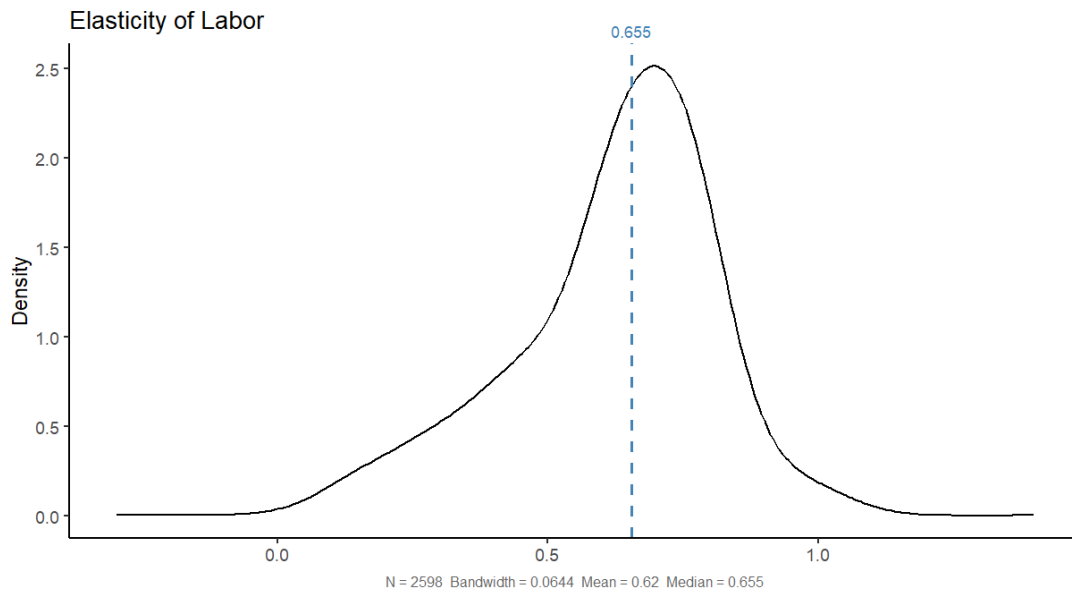
Notes: $\overline{ES}$ denotes the sample mean of the pointwise elasticity of scale estimates, while $\widehat{ES}(\bar{x})$ denotes the elasticity of scale evaluated at the sample mean of the input variables. Probit and logit specifications yield identical estimates and are therefore reported jointly in a single row.

Figure 3.5: Empirical distribution of the capital elasticity: PLM-IV-NP estimates



Note: The figure presents the kernel density estimate of the firm-level capital elasticity $\hat{\beta}_k$, obtained from the semiparametric partially linear model with endogenous process innovation. The dashed vertical line indicates the median value, as reported on the horizontal axis. The sample mean is $\bar{\beta}_k = 0.193$. The estimation is based on a sample of $N = 2,598$ observations, and the bandwidth is selected according to Silverman's rule of thumb.

Figure 3.6: Empirical distribution of the labor elasticity: PLM-IV-NP estimates



Note: The figure presents the kernel density estimate of the firm-level labour elasticity $\hat{\beta}_l$, obtained from the semiparametric partially linear model with endogenous process innovation. The dashed vertical line indicates the median value, as reported on the horizontal axis. The sample mean is $\bar{\beta}_l = 0.620$. The estimation is based on a sample of $N = 2,598$ observations, and the bandwidth is selected according to Silverman's rule of thumb.

3.5 Conclusions

We propose a novel estimator for partially linear models in which a dummy endogenous variable enters the model parametrically. The framework departs from the standard DEVM by allowing exogenous covariates to enter through an unknown smooth function, while preserving a constant structural shift associated with the endogenous dummy variable. Monte Carlo evidence indicates that the proposed estimator exhibits satisfactory finite-sample performance. The empirical application to the knowledge capital production function further highlights the practical relevance of the approach: by relaxing parametric functional form assumptions and explicitly accounting for the endogeneity of the dummy variable, the method delivers estimates of process innovation that are consistent with those obtained from standard parametric specifications, while simultaneously uncovering a richer and more informative pattern of input elasticities that cannot be captured within a purely parametric framework.

Chapter 4

Mixed-type Endogenous Variables in Partially Linear Regression Model

Abstract

We propose a novel estimator for partially linear models with mixed-type endogenous variables, in which both a continuous and a binary endogenous regressor enter the model simultaneously. The methodology integrates a control function approach to address the endogeneity of the continuous regressor with an instrumental variable strategy that employs fitted probabilities as generated instruments for the binary endogenous regressor. Monte Carlo simulations are conducted to evaluate the finite-sample performance of the estimator under varying degrees of endogeneity. The results indicate that the proposed procedure performs satisfactorily across different scenarios. Finally, the empirical application to a knowledge capital production function, based on firm-level data, illustrates the practical relevance of the approach by jointly addressing the endogeneity of labor and process innovation.

4.1 Introduction

Estimating the returns to innovative activity at the firm level constitutes a central concern in the economics of technological change (Pakes & Griliches 1984; B. H. Hall, Mairesse, & Mohnen 2010; Acemoglu, Akcigit, et al. 2018). Empirical research in this area has frequently relied on the knowledge capital production function framework developed by Pakes & Griliches (1984) & Griliches (1998), which postulates a causal relationship between unobservable knowledge capital and observable indicators of innovation inputs such as R&D expenditures, innovation outputs such as patents, and firm-level performance. Building on this framework, Crépon et al. (1998) introduced a structural model in which research investment determines innovation output, which in turn determines productivity, along with a method for correcting the selectivity and endogeneity inherent in the specification. Known as the CDM model, this framework has become the standard reference for firm-level empirical analysis of the innovation–productivity relationship (Löf et al. 2017).

The growing availability of harmonised Community Innovation Surveys (CIS) (Mairesse & Mohnen 2010; Mohnen & B. H. Hall 2013) has facilitated the construction of direct binary measures of innovation output as an alternative to R&D-based measures (B. H. Hall & Mairesse 1995; Musolesi & Huiban 2010; Mairesse & Mohnen 2010; Mohnen & B. H. Hall 2013; Antonioli et al. 2021). These binary indicators have emerged as one of the most widely adopted alternatives in the empirical literature (Huergo & Jaumandreu 2004; Griffith et al. 2006; Masso & Vahter 2008; Chudnovsky et al. 2006; Parisi et al. 2006; Duguet 2006), as they allow direct comparability across studies and survey waves (B. H. Hall, Lotti, et al. 2009; B. H. Hall 2011; Mohnen & B. H. Hall 2013).

The inclusion of a binary measure of innovation within the production function

framework provides a coherent means of capturing the effect of technical change on the production process, following the econometric approach of Duguet (2006), Parisi et al. (2006), & Musolesi & Huiban (2010). This specification is consistent with the *Hicksian* notion of neutral technical change (Hicks 1932), according to which innovation shifts the production frontier proportionally without altering the relative marginal products of inputs; for a detailed discussion, see Antonioli et al. (2021). On this basis, we specify the following production function

$$y_i = \beta_0 + \beta_K k_i + \beta_L \ell_i + \beta_{IP} IP_i + \varepsilon_i, \quad (4.1)$$

where y_i denotes log value added, k_i and ℓ_i denote log capital and log labour, IP_i is a binary indicator of process innovation, and sectoral dummies are included throughout. The inclusion of a dummy endogenous regressor in this specification places the present framework within the dummy endogenous variable model (DEVM) of J. J. Heckman (1978), itself a special case of the hybrid model (Maddala 1983).

With regard to the choice of innovation variable, this chapter focuses on process innovation. A growing body of work (Jaumandreu & Mairesse 2010; Jaumandreu & Mairesse 2017) argues that process innovation raises productivity primarily through reductions in production costs, whereas product innovation operates mainly through shifts in the demand function. This distinction is further reinforced by the literature on technology and employment, which emphasizes the asymmetric labour-market effects associated with these two forms of innovation (Vivarelli 2014; Harrison et al. 2014; Calvino & Virgillito 2018). In particular, the notion of technological unemployment is more closely linked to process innovation than to product innovation, since the latter tends to exert a labour-friendly effect through the creation of new markets and the expansion of demand (Bogliacino et al. 2012).

Regarding the choice of functional form, the literature has traditionally relied on the Cobb-Douglas specification, which imposes strong *a priori* restrictions, most notably unitary elasticities of substitution and input separability (Cobb & Douglas 1928). Subsequent contributions have sought to relax these restrictions by allowing for more flexible parametric representations (Christensen et al. 1973; Gallant 1981; Ivaldi et al. 1996). These approaches, however, continue to impose a fully specified functional form. By contrast, the nonparametric and semiparametric literature proceeds one step further by leaving the input-output relationship unspecified (Henderson & Russell 2005; Henderson & Kumbhakar 2006; Kumbhakar, Park, et al. 2007; Kumbhakar, Sun, et al. 2016; Ma et al. 2020). Motivated by this literature, we consider the following partially linear specification

$$y_i = m(k_i, \ell_i) + \beta IP_i + \epsilon_i, \quad (4.2)$$

where $m(\cdot)$ is an unknown smooth function of the continuous inputs and the remaining terms are defined as in equation (4.1).

To address the endogeneity of the continuous labour regressor, we adopt the CF approach (Newey, Powell, & Vella 1999; Pinkse 2000; Su & Ullah 2008; Martins-Filho & Yao 2012; Delgado & Parmeter 2014), which controls for the dependence between the structural disturbance and the endogenous regressor by augmenting the structural equation with nonparametric residuals obtained from a reduced-form first stage. To address the endogeneity of the binary innovation indicator entering parametrically, we employ an IV estimator (Li & Stengos 1996; Li & J. S. Racine 2007), using fitted probabilities from a binary response model as generated instruments (Wooldridge 2003; Wooldridge 2010).

As regards the endogeneity of labor, the cross-sectional nature of the data precludes the use of structural identification strategies that exploit panel variation (Olley & Pakes 1996; D. A. Akerberg et al. 2015). Although input and out-

put prices may serve as instruments in this setting (Griliches & Mairesse 1995; D. Akerberg et al. 2007; D. A. Akerberg et al. 2015), such variables are often unavailable or display insufficient variation, and their validity requires price dispersion to be driven by exogenous market conditions—an assumption more plausibly satisfied by lagged prices (Doraszelski & Jaumandreu 2013). Since labor is the more flexible input, correcting for its endogeneity is the more pressing concern, whereas capital is treated as predetermined (Doraszelski & Jaumandreu 2013). Moreover, transmission bias appears to be empirically limited in value-added specifications (D. Akerberg et al. 2007), while data constraints preclude the construction of a reliable measure of the user cost of capital (Jaumandreu & Mairesse 2006). We therefore instrument labor using lagged wages and treat capital as exogenous.

The empirical analysis relies on firm-level data from the tenth Unicredit–Capitalia survey merged with balance-sheets from AIDA or the Chambers of Commerce (Antonioli et al. 2021). Value added, capital, and labor are in natural logarithms for 2006. Process innovation is defined as a binary variable equal to one if a firm reports at least one process innovation between 2004 and 2006. Manufacturing firms are classified into 22 sectors according to the two-digit ATECO2002 classification. IVs for the innovation indicator are selected through the multi-step procedure proposed by Antonioli et al. (2021), which combines knowledge inputs and absorptive-capacity factors. The final set of instruments consists of *BetterProd* (investment in product improvement), *FinanIncent* (receipt of fiscal or financial incentives), and *MarkPenEU15* (market penetration in the EU15 market).

The remainder of the chapter is organized as follows. Section 4.2 presents the general econometric specification and develops the consistent estimation of partially linear models with mixed endogenous regressors. Section 4.3 reports Monte

Carlo simulations designed to assess the finite-sample properties of the proposed estimator. Section 4.4 applies the methodology to the estimation of the knowledge production function in the presence of endogenous labor and process innovation. Finally, Section 4.5 concludes.

4.2 Econometric Model and Estimation Procedure

We begin by considering the estimation of the following partially linear production function

$$y_i = m(k_i, \ell_i) + \sum_{j=1}^J \beta_j S_{ji} + \beta_{IP} IP_i + \epsilon_i, \quad (4.3)$$

where k_i and ℓ_i denote the logarithms of capital and labor, respectively, IP_i is a dummy variable indicating the presence of process innovation, and S_{ji} are sectoral dummies. Let G denote the total number of sectors and set $J = G - 1$, so that one sector is omitted and serves as the reference category. Finally, ϵ_i denotes the error term.

The model incorporates two endogenous variables. Labor, ℓ_i , is continuous and enters the specification nonparametrically, whereas the process innovation dummy, IP_i , enters parametrically and generates a structural shift in the production function.

Before turning to the estimation procedure, it is useful to introduce the kernel functions employed to smooth over both continuous and categorical variables. Let $Z_i = (Z_{1i}, Z_{2i})$ denote a $(q_z \times 1)$ vector of instruments, where Z_{1i} is a $(q_{z_1} \times 1)$ vector of instruments for W_i^c and Z_{2i} is a $(q_{z_2} \times 1)$ vector of instruments for W_i^d , with $q_z = (q_{z_1} + q_{z_2})$. For r -th component of Z_{2i} we assume that Z_{2ir} may take $(c_r - 1)$ distinct values, that is, $Z_{2ir} \in \{0, 1, \dots, c_r - 1\}$ for $r = 1, \dots, q_{z_2}$. Furthermore, let $z = (z_1, z_2) \in \mathbb{R}^{q_z}$, let $f(z) = f(z_1, z_2)$ denote the joint density function of (Z_{1i}, Z_{2i}) ,

and define $\mathcal{D} = \prod_{r=1}^{q_{z_2}} \{0, 1, \dots, c_r - 1\}$ as the support of Z_{2i} .

For the continuous random variables, Z_{1i} , we define the product kernel as

$$K_h(Z_{1i}, z_1) = \prod_{s=1}^{q_{z_1}} \frac{1}{h_s} k\left(\frac{Z_{1is} - z_{1s}}{h_s}\right), \quad (4.4)$$

where $k(\cdot)$ is the univariate kernel function and $h = (h_1, \dots, h_{q_{z_1}})$ is the vector of bandwidths.

For the discrete regressors, we focus on the case in which no natural ordering is present, as in the binary cases, and define

$$\ell(Z_{2ir}, z_{2r}, \lambda) = \begin{cases} 1 & \text{if } Z_{2ir} = z_{2r} \\ \lambda & \text{if } Z_{2ir} \neq z_{2r}, \end{cases} \quad (4.5)$$

where λ is a smoothing parameter taking values in the interval $[0, 1]$. Thus, $\ell(Z_{2i}, z_2, 0)$ reduces to an indicator function when $\lambda = 0$, whereas $\ell(Z_{2i}, z_2, \lambda)$ becomes a uniform weighting function when $\lambda = 1$. Let $\mathbb{1}(Z_{2ir} \neq z_{2r})$ denote the indicator function taking value one whenever $(Z_{2ir} \neq z_{2r})$, and zero otherwise. We further define $d_{Z_i, z} = \sum_{r=1}^{q_{z_2}} \mathbb{1}(Z_{2ir} \neq z_{2r})$ as the number of disagreeing components. The product kernel for the discrete random variables may then be written as

$$L(Z_{2i}, z_2, \lambda) = \prod_{r=1}^{q_{z_2}} \ell(Z_{2ir}, z_{2r}, \lambda) = \lambda^{d_{Z_i, z}}, \quad (4.6)$$

where $\lambda = (\lambda_1, \dots, \lambda_{q_{z_2}})$ is the vector of smoothing parameters.

Let $\phi = (h, \lambda) = (h_1, \dots, h_{q_{z_1}}, \lambda_1, \dots, \lambda_{q_{z_2}})$ collect the smoothing parameters associated with the continuous and discrete kernel components. The joint kernel weight function for mixed continuous and discrete variables is then defined as:

$$\mathcal{W}_\phi(Z_i, z) = \prod_{s=1}^{q_{z_1}} \frac{1}{h_s} k\left(\frac{Z_{1is} - z_{1s}}{h_s}\right) \prod_{r=1}^{q_{z_2}} \ell(Z_{2ir}, z_{2r}, \lambda) = K_h(Z_{1i}, z_1) L(Z_{2i}, z_2, \lambda). \quad (4.7)$$

4.2.1 Estimation Method

We consider the following triangular simultaneous equations model,

$$\begin{cases} Y_i = m(X_{1i}, W_i^c) + \beta_1 X_{2i}^d + \beta_2 W_i^d + \epsilon_i, \\ W_i^c = g(Z_{1i}) + u_i, \end{cases} \quad E(\epsilon_i | Z_{1i}) = 0, \quad E(\epsilon_i | Z_{1i}, u_i) = E(\epsilon_i | u_i), \quad (4.8)$$

where X_{1i} is a $(q_{x_1} \times 1)$ vector of continuous exogenous variables entering the model nonparametrically, such as capital k , and X_{2i}^d is a discrete exogenous variable entering parametrically, such as sectorial dummies. For expositional simplicity, we treat X_{2i}^d as binary, although the extension to a vector-valued case is straightforward. Moreover, W_i^c denotes a vector of endogenous continuous regressors, such as labor ℓ , while W_i^d denotes the dummy endogenous regressor, namely process innovation, IP . In addition, $Z_{1i} = (X_{1i}, X_{3i})$ and $Z_{2i} = (X_{2i}^d, X_{4i})$ are $(q_{z_1} \times 1)$ and $(q_{z_2} \times 1)$ vectors comprising included exogenous regressors, (X_{1i}, X_{2i}^d) , together with excluded instruments, (X_{3i}, X_{4i}) , for W_i^c and W_i^d , respectively. These instruments may themselves be continuous or discrete. The functions $m(\cdot)$ and $g(\cdot)$ are unknown smooth functions, and u_i is the idiosyncratic error term.

Direct estimation of (4.8) using standard procedures such as Robinson (1988) yields inconsistent estimates of the parameters of interest because of the endogeneity of both W_i^c and W_i^d . More precisely, ignoring the endogeneity of either regressor invalidates the moment conditions underlying the Robinson-type partialling-out step, since the transformed errors remain correlated with both W_i^c and W_i^d , given that $\mathbb{E}[\epsilon | W_i^c] \neq 0$ and $\mathbb{E}[\epsilon | W_i^d] \neq 0$. This contamination destroys the orthogonality conditions required for the consistent estimation of both the parametric and nonparametric components.

To address endogeneity in the nonparametric component, for given values of β_1 and β_2 , we rewrite (4.8) as the following fully nonparametric model

$$Y_i^* = m(X_{1i}, W_i^c) + \epsilon_i, \quad (4.9)$$

where $Y_i^* = Y_i - \beta_1 X_{2i}^d - \beta_2 W_i^d$.

A number of methodologies have been proposed in the literature to handle endogeneity, including IVs, generalized method of moments (GMM), maximum likelihood estimation, and the CF approach. In this study, we adopt the CF approach, since it provides a natural framework for dealing with nonlinear relationships in the presence of endogeneity (Newey, Powell, & Vella 1999; Su & Ullah 2008; Delgado & Parmeter 2014). More specifically, the CF method mitigates endogeneity by incorporating a CF into the structural equation, thereby capturing the dependence structure of the error term. In contrast to IV methods, which generally require strong exclusion restrictions in nonlinear settings, the CF approach may offer greater robustness, particularly when the identification of valid instruments is difficult (J. J. Heckman & Vytlacil 2005).

Taking conditional expectations on both sides of the model and using the assumptions stated above, it follows that the conditional mean no longer coincides with the structural object of interest. In particular,

$$\begin{aligned}\mathbb{E}[Y_i^* | X_{1i}, W_i^c, Z_{1i}, u_i] &= m(X_{1i}, W_i^c) + \mathbb{E}[\epsilon_i | X_{1i}, W_i^c, Z_{1i}, u_i] \\ &= m(X_{1i}, W_i^c) + r(u_i),\end{aligned}\tag{4.10}$$

since $\mathbb{E}[\epsilon_i | X_{1i}, W_i^c, Z_{1i}, u_i] = \mathbb{E}[\epsilon_i | u_i] \equiv r(u_i)$, where $r(u_i)$ is an unknown function. Thus, in order to control for the endogeneity of W_i^c , we propose estimating the following partially linear model,

$$Y_i = m_0(X_{1i}, W_i^c, u_i) + \beta_1 X_{2i}^d + \beta_2 W_i^d + \eta_i.\tag{4.11}$$

where $m_0(X_{1i}, W_i^c, u_i) = m(X_{1i}, W_i^c) + r(u_i)$, and $\eta_i = \epsilon_i - r(u_i)$ is the new error term. It should be noted that the conditional mean independent assumption imposed in (4.8) is crucial for ensuring the additive separability of the unknown function $m(\cdot)$. In particular, this assumption guarantees that the endogenous compo-

nent enters exclusively through $r(u_i)$, separately from the structural relationship $m(X_{1i}, W_i^c)$, thereby allowing the structural function to be identified and estimated separately of the nuisance CF.

Nevertheless, estimation of $m(\cdot)$ raises a fundamental identification problem, since the function $m_0(X_{1i}, W_i^c, u_i)$ admits an additive decomposition in which $m(X_{1i}, W_i^c)$ represents the structural function of interest, while $r(u_i)$ plays the role of the CF correction term (Newey, Powell, & Vella 1999; Su & Ullah 2008; Delgado & Parmeter 2014; Henderson & Parmeter 2015).

4.2.2 Estimation Procedure

To obtain consistent estimates in (4.11), we propose a new estimation procedure consisting of three sequential stages, designed to address both sources of endogeneity in a systematic way.

The final model to be estimated may be rewritten in compact form as

$$Y_i = m_0(X_{1i}, W_i^c, u_i) + \beta V_i + \eta_i \quad (4.12)$$

where $V_i = (X_{2i}^d, W_i^d)^\top$ collects the regressors entering parametrically, $\beta = (\beta_1, \beta_2)$ is the vector of associated coefficients, and η_i is the new error term.

First Stage

When multiple exogenous variables are available in Z_{2i} to instrument for W_i^d , standard IV estimation is no longer directly applicable, since it requires a single instrument. A natural alternative is two-stage least squares (2SLS), which accommodates multiple instruments by projecting W_i^d onto the full instrument set Z_{2i} in the first stage. However, when W_i^d is binary, this linear projection gives rise to the forbidden regression problem (Wooldridge 2010): the first-stage fitted values do not recover the conditional expectation $E(W_i^d | Z_{2i})$, and their direct use as in-

struments in the second stage does not yield consistent estimates of the structural parameters.

To overcome this issue, J. D. Angrist (2001), Wooldridge (2003), & Wooldridge (2010) propose replacing the linear first-stage projection with the conditional expectation $\mathbb{E}[W_i^d | Z_{2i}] = P(W_i^d = 1 | Z_{2i}) \equiv p(Z_{2i})$, which aggregates all relevant exogenous variation in Z_{2i} into a single generated instrument while preserving the binary nature of W_i^d .

Accordingly, the first stage serves a dual purpose. First, it generates fitted probabilities that are subsequently used as instruments in the second stage to address the endogeneity of the binary regressor W_i^d . Second, it produces consistent estimates of the reduced-form residuals required to control for the endogeneity of the continuous regressor W_i^c within the CF framework. The function $p(\cdot)$ captures the variation in W_i^d induced by the instruments and is therefore orthogonal to the structural disturbance.

Following Wooldridge (2010), fitted values from any first-stage model that consistently estimates $p(Z_{2i}) = E(W_i^d | Z_{2i})$ may be used as instruments for W_i^d , provided that $p(Z_{2i})$ is not perfectly collinear with the included exogenous covariates (X_{2i}^d, X_{4i}) . This condition is automatically satisfied whenever $p(Z_{2i})$ is a nonlinear function of Z_{2i} , as is generically the case under correctly specified binary response models. Wooldridge (2010) further shows that consistency on the second-stage estimator does not require correct parametric specification of the first stage, but only that $p(Z_{2i})$ remains a relevant instrument after linearly partialling out the included exogenous variables. In the present semiparametric setting, this requirement applies after the Robinson-type partialling-out step that removes the nonparametric component $m_0(X_{1i}, W_i^c, u_i)$, that is, the residual variation in $p(Z_{2i})$, orthogonal to the partialled-out regressors, must remain correlated with W_i^d . This condition holds whenever Z_{2i} contains at least one variable that shifts

$E(W_i^d|Z_{2i})$, which is precisely the standard instrument relevance condition.

Although consistency is robust to first-stage misspecification, parametric binary response models such as probit and logit impose functional form restrictions on $p(Z_{2i})$ that may be violated in practice, thereby introducing bias in the generated instruments and propagating misspecification into the second stage. The choice of first-stage estimator therefore reflects a bias-variance trade-off: parametric models are efficient under correct specification but potentially severely biased under misspecification, whereas semiparametric and nonparametric alternatives sacrifice some efficiency in order to reduce such bias. To guard against misspecification, we estimate $p(Z_{2i})$ using methods that relax parametric restrictions. In particular, we consider the semiparametric single-index of Klein & Spady (1993), which preserves a parametric index structure while estimating the link function nonparametrically and attains the semiparametric efficiency bound for binary response models, and the local likelihood logit estimator of Frölich (2006), which estimates $p(Z_{2i})$ fully nonparametrically through kernel-weighted local maximum likelihood, imposing no parametric restrictions on the conditional expectation $E(W_i^d|Z_{2i})$, albeit at the cost of slower convergence rates. The second objective of this stage is to recover the latent disturbance u_i from the reduced-form equation for the continuous endogenous regressor, in order to control for the endogeneity of W_i^c . Since u_i is unobserved, the structural representation in (4.12) is not directly estimable.

We assume that the endogenous regressor satisfies

$$W_i^c = g(Z_{1i}) + u_i. \quad (4.13)$$

where $g(\cdot)$ may be estimated using local linear regression techniques (Stone 1977; Cleveland 1979; Fan & Gijbels 1992). The corresponding nonparametric residu-

als are then given by

$$\widehat{u}_i = W_i^c - \widehat{g}(Z_{1i}). \quad (4.14)$$

where $\widehat{g}(\cdot)$ denotes the corresponding nonparametric estimator of $g(\cdot)$.

Second Stage

To control for the continuous endogenous variables in (4.12), we use the nonparametric residuals obtained in the first stage and rewrite the model as

$$Y_i = m_0(X_{1i}, W_i^c, \widehat{u}_i) + \beta^\top V_i + e_i, \quad (4.15)$$

where $e_i = \eta_i + \{r(u_i) - r(\widehat{u}_i)\}$ is the new error term.

To address the dummy endogenous variable entering parametrically through $V_i = (X_{2i}^d, W_i^d)$, we adopt the IV estimator (Li & Stengos 1996; Li & J. S. Racine 2007) using fitted probabilities as generated instruments. Applying the transformation proposed by Robinson (1988), we partial out the nonparametric component by defining $\widetilde{Y}_i = Y_i - E[Y_i | \chi_i]$ and $\widetilde{V}_i = V_i - E[V_i | \chi_i]$, where $\chi_i = (X_{1i}, W_i^c, \widehat{u}_i)$, yielding

$$\widetilde{Y}_i = \beta^\top \widetilde{V}_i + e_i. \quad (4.16)$$

Standard approaches remain inconsistent because $E[\eta_i | W_i^d] = E[e_i | W_i^d] \neq 0$, thus violating the orthogonality required for consistent estimation. Since W_i^d is binary, both standard IV and linear 2SLS are affected by the forbidden regression problem. We therefore use $\widehat{p}_i = \widehat{p}(Z_{2i})$ as the instrument. Under a correctly specified binary first-stage model, $\widehat{p}_i$ approximates the optimal instrument $E(W_i^d | Z_{2i})$, thereby minimizing the asymptotic variance of the IV estimator and dominating linear 2SLS (Wooldridge 2010). Defining $\Theta_i = (X_{2i}^d, \widehat{p}_i)$ and $\beta_{\widehat{p}} = (\beta_1, \beta_2)$, the feasible estimator is given by

$$\widehat{\beta}_{\widehat{p}} = \left(\sum_{i=1}^N \widehat{\Theta}_i \widehat{\widetilde{V}}_i^\top \widehat{f}^2(\chi_i) \right)^{-1} \sum_{i=1}^N \widehat{\Theta}_i \widehat{\widetilde{Y}}_i^\top \widehat{f}^2(\chi_i), \quad (4.17)$$

where the conditional expectations and the marginal density $f(\cdot)$ are estimated using Nadaraya-Watson kernel regressions and kernel density estimator.

Third Stage

To estimate the unknown smooth function $m(\cdot)$, we use the second-stage results to rewrite (4.15) as the following fully nonparametric model

$$Y_{0i} = m_0(X_{1i}, W_i^c, \widehat{u}_i) + \omega_i, \quad (4.18)$$

where $Y_{0i} = Y_i - \widehat{\beta}^\top V_i$ and $\omega_i = e_i + \{\beta - \widehat{\beta}\}^\top V_i$.

To estimate $m_0(\cdot)$, we apply a local linear regression to equation (4.18) (Fan & Gijbels 1992). Let $\chi_i = (X_{1i}, W_i^c, \widehat{u}_i)$ and $\chi_0 = (x_1, w^c, \widehat{u})$ denote an interior point in the support of χ_i . The local linear estimator is then given by

$$\begin{aligned} \widehat{m}_0(x_1, w^c, \widehat{u}) = & e_1^\top \left[\sum_{i=1}^N K_{h_3}(\chi_i, \chi_0) \begin{pmatrix} 1 & (\chi_i - \chi_0)^\top \\ (\chi_i - \chi_0) & (\chi_i - \chi_0)(\chi_i - \chi_0)^\top \end{pmatrix} \right]^{-1} \\ & \times \sum_{i=1}^N K_{h_3}(\chi_i, \chi_0) \begin{pmatrix} 1 \\ (\chi_i - \chi_0) \end{pmatrix} Y_{0i}, \end{aligned} \quad (4.19)$$

where h_3 is the bandwidth parameter associated with this stage.

To recover the marginal effect $m(x_1, w^c)$, we integrate out the generated regressor $\widehat{u}_i$ using the marginal integration approach independently introduced by Tjøstheim & Auestad (1994) and Linton & Nielsen (1995),

$$\widehat{m}(x_1, w^c) = \int \widehat{m}_0(x_1, w^c, \widehat{u}) \omega(\widehat{u}) d\widehat{u}, \quad (4.20)$$

where $\omega(\cdot)$ is a weighting function satisfying $\int_{\mathbb{W}} \omega(u) du = 1$, and $\mathbb{W}$ denotes the compact support of u_i .

In principle, the integral in (4.20) may be evaluated numerically using methods such as Simpson's rule or the trapezoidal rule. However, if $\mathcal{S}$ evaluation points are employed in each dimension of $\widehat{u}_i$, the calculation of each point in (x_1, w^c)

requires $\mathcal{S}^{(q_{x_1}+q_w)}$ weighted least squares problems, each involving $O(N)$ operations. This computational burden may become prohibitive in higher-dimensional settings.

To address this issue, we adopt the computationally efficient marginal integration strategy proposed by Qian & Wang (2012) and Delgado & Parmeter (2014). The basic idea is to generate *i.i.d.* draws $(\widehat{u}_\ell^*, \ell = 1, \dots, N)$ from $\omega(\cdot)$ and construct $\widehat{m}_{MC}(x_1, w^c) = \frac{1}{N} \sum_{\ell=1}^N \widehat{m}_0(x_1, w^c, \widehat{u}_\ell^*)$, which provides a good approximation to $\widehat{m}(\cdot)$ for sufficiently large N . When $\omega(\cdot)$ is chosen as the empirical density of $\widehat{u}_i$, this simplifies to the sample average

$$\widehat{m}(x_1, w^c) = \frac{1}{N} \sum_{i=1}^N \widehat{m}_0(x_1, w^c, \widehat{u}_i), \quad (4.21)$$

where $\widehat{m}_0(\cdot)$ is obtained from the local linear regression in (4.19) (Stone 1977; Cleveland 1979; Fan & Gijbels 1992). Under $E(\varpi_i) = 0$, this estimator shares the same asymptotic properties as (4.20).

4.2.3 Bandwidth Selection Procedure

Several bandwidth selection procedures are available in the literature, including Rule-of-Thumb (RoT) methods, plug-in procedures, cross-validation (CV), and bootstrap techniques. Among these, CV is often particularly effective in local linear kernel estimation with discrete regressors (P. Hall, J. Racine, et al. 2004). Following Martins-Filho & Yao (2012), we adopt the RoT bandwidth selection method in all three stages of the estimation process. In triangular simultaneous models, the asymptotic distribution of local polynomial estimators is unaffected by the first-stage bandwidth (Su & Ullah 2008), and an optimal first-stage bandwidth does not necessarily imply optimal performance of the resulting nonparametric estimator (Henderson & Parmeter 2015).

The formal theory underlying RoT bandwidth selection is available only for con-

tinuous covariates. In the case of continuous variables, the bandwidth depends on the standard deviation of the covariate, whereas no analogous theoretical result exists for discrete variables. Following Henderson & Parmeter (2015), we therefore employ pilot bandwidths based on sample size for the discrete case. The optimal bandwidth rates are given by $h \sim N^{-\frac{1}{4+q_c}}$, and $\lambda \sim N^{-\frac{2}{4+q_c}}$, where q_c denotes the number of continuous regressors. For continuous covariates, the bandwidth is additionally scaled by the variability of the covariate, yielding $h = c\sigma_x N^{-\frac{1}{4+q_c}}$, where c is a scale factor and σ_x denotes the standard deviation of the covariate. For discrete covariates, we adopt the practical sample-size-based approach: $\lambda = cN^{-\frac{2}{4+q_c}}$.

4.3 Monte Carlo Simulations

To evaluate the performance of the estimators proposed in this chapter, we conduct a Monte Carlo simulation based on the following data-generating processes (DGPs),

$$\text{DGP1 : } Y_i = -0.5 X_{1i} + \sin(2 W_{1i}^c) + \frac{4}{1 + \exp(-3(X_{1i} + W_{1i}^c - 1))} + \beta_1 X_{2i}^d + \beta_2 W_i^d + \epsilon_i, \quad (4.22)$$

$$\text{DGP2 : } Y_i = X_{1i} + \sin(W_{1i}^c) + \beta_1 X_{2i}^d + \beta_2 W_i^d + \epsilon_i. \quad (4.23)$$

where $\beta_1 = \beta_2 = 1$ and the covariates are generated as random variables such that $X_{1i} \sim i.i.d. \mathcal{U}(-5, 5)$, while $X_{2i}^d \in \{0, 1\}$ with $Pr[X_{2i}^d = 1] = 0.5$. To induce correlation between W_{1i}^c and the disturbance term, we specify

$$W_{1i}^c = c_1 Z_{1i}^c + \sin(Z_{1i}^c) + Z_{2i}^c + (Z_{2i}^c)^2 + u_i, \quad (4.24)$$

with c_1 is a constant governing instrument strength, with $c_1 = 1$ corresponding to strong instruments and $(0 < c_1 < 1)$ to weaker ones, and where $Z_{1i}^c, Z_{2i}^c \sim i.i.d.$

$\mathcal{U}(-5, 5)$.

To generate correlation between W_i^d and the disturbance term, we define

$$W_i^{d*} = c_0 + c_1 Z_{1i}^d + (Z_{1i}^d)^2 + \sin(Z_{2i}^d) + v_i, \quad (4.25)$$

$$W_i^d = \mathbb{1}(W_i^{d*} > 0), \quad (4.26)$$

where c_0 is a threshold-shifting constant in the latent index, calibrated so that $\Pr(W_i^d = 1 \mid \mathbf{Z}_i^d) \approx \Pr(W_i^d = 0 \mid \mathbf{Z}_i^d) \approx 0.5$, and c_1 plays the same role as above. In addition, $Z_{1i}^d, Z_{2i}^d \sim i.i.d. \mathcal{U}(-5, 5)$, and $\mathbb{1}(\cdot)$ denotes the indicator function.

We further assume that ϵ_i , u_i , and v_i are jointly normally distributed according to

$$\begin{pmatrix} \epsilon_i \\ u_i \\ v_i \end{pmatrix} \sim i.i.d. \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta_1 & \theta_2 \\ \theta_1 & 1 & 0 \\ \theta_2 & 0 & 1 \end{pmatrix} \right), \quad (4.27)$$

where $\theta_1 = \theta_2 = (0.2, 0.5, 0.8)$ govern the degree of endogeneity.

The number of Monte Carlo replications is set to 1,000. We employ the Epanechnikov kernel function $k(u) = 0.75(1 - u^2)\mathbb{1}\{|u| \leq 1\}$ and consider sample sizes $N \in \{100, 200, 400\}$. To evaluate the performance of the estimator proposed for β in (4.17), we report bias, Standard deviation (Sd), and Root Mean Squared Error (RMSE). For the nonparametric estimator proposed for $m(\cdot)$ in (4.19), performance is assessed using the RMSE. These measures are computed as

$$RMSE_{\beta} = \left[\frac{1}{R} \sum_{r=1}^R (\widehat{\beta}_r - \beta)^2 \right]^{\frac{1}{2}} \quad (4.28)$$

$$RMSE_m = \left[\frac{1}{N} \sum_{i=1}^N (\widehat{m}_r(x_{1i}, w_i^c) - m_r(x_{1i}, w_i^c))^2 \right]^{\frac{1}{2}}, \quad (4.29)$$

where the subscript r indexes the r -th replication out of a total of R . To ensure robustness and reliability, we restrict the analysis to the 10%-90% quantile range of the simulation outcomes, thereby excluding extreme values, as recommended by Martins-Filho & Yao (2012) and Delgado & Parmeter (2014), among others.

Table 4.1: Finite sample performance of $\widehat{\beta}$: single index first stage

θ	N	DGP1			DGP2		
		<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>	<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>
0.2	100	2.3213	2.2316	3.22	2.2740	2.0327	3.0501
	200	0.7987	0.5943	0.9956	0.7888	0.5016	0.9348
	400	0.3599	0.3150	0.4783	0.3539	0.2400	0.4276
0.5	100	2.3605	2.1872	3.2180	2.3314	1.9521	3.0408
	200	0.8182	0.5791	1.0024	0.8096	0.4844	0.9435
	400	0.3704	0.3073	0.4813	0.3605	0.2296	0.4274
0.8	100	2.4569	2.3229	3.3811	2.4116	1.9381	3.0938
	200	0.8375	0.5676	1.0117	0.8299	0.4599	0.9488
	400	0.3787	0.2926	0.4786	0.3633	0.2148	0.4221

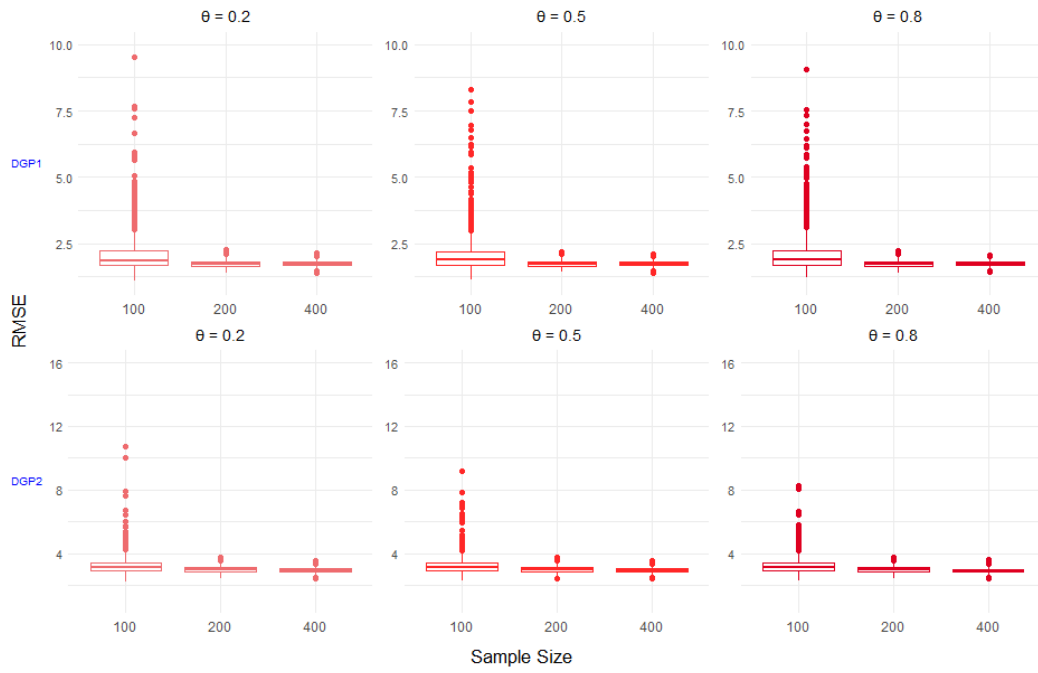
Note: The table reports average Bias, standard deviation (Sd), and RMSE of $\widehat{\gamma}$ over 1,000 replications for the three-stage estimator under DGP1 and DGP2.

Table 4.2: Finite sample performance of $\widehat{\beta}$: local logit first stage

θ	N	DGP1			DGP2		
		<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>	<i>Bias</i>	<i>Sd</i>	<i>RMSE</i>
0.2	100	2.3732	2.2254	3.2533	2.3190	2.0405	3.0889
	200	0.8214	0.5781	1.0044	0.8111	0.4834	0.9442
	400	0.3733	0.3109	0.4858	0.3677	0.2373	0.4376
0.5	100	2.4633	2.2121	3.3108	2.4229	1.9962	3.1393
	200	0.8763	0.5672	1.0439	0.8659	0.4684	0.9845
	400	0.4031	0.3033	0.5044	0.3949	0.2249	0.4545
0.8	100	2.6039	2.4581	3.5809	2.5494	2.0277	3.2574
	200	0.9287	0.5610	1.0850	0.9130	0.4529	1.0192
	400	0.4249	0.2862	0.5123	0.4089	0.2088	0.4591

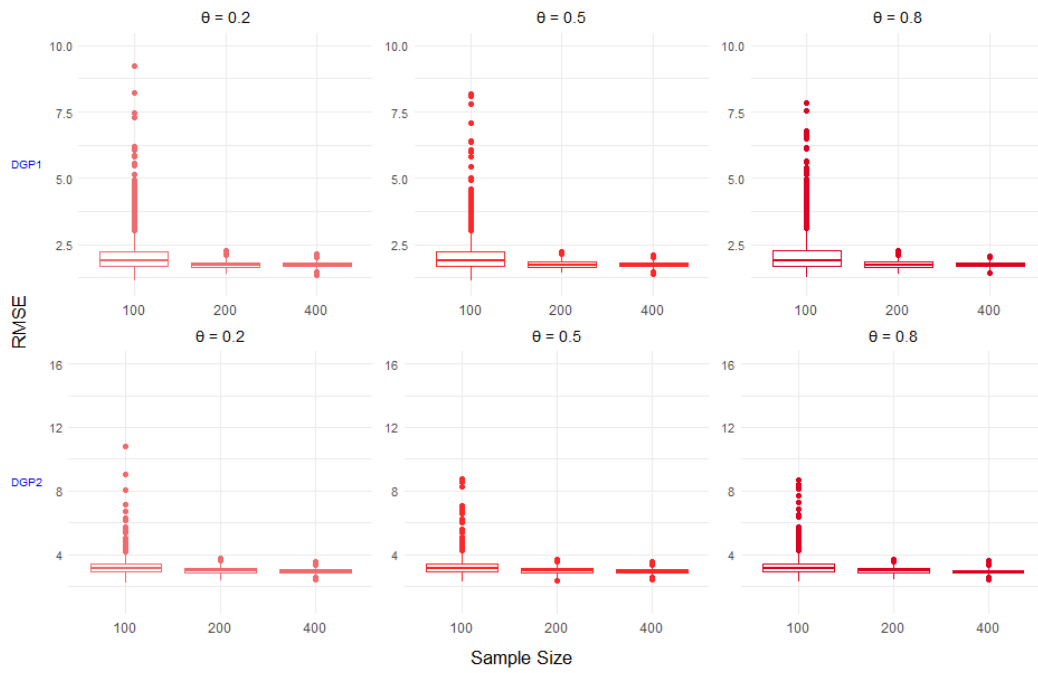
Note: The table reports average Bias, standard deviation (Sd), and RMSE of $\widehat{\gamma}$ over 1,000 replications for the three-stage estimator under DGP1 and DGP2.

Figure 4.1: Finite sample performance of $\widehat{m}(\cdot)$: single index first stage



Note: Boxplots of the RMSE values of the nonparametric estimator of $m(\cdot)$ in 1,000 replications.

Figure 4.2: Finite sample properties of $\widehat{m}(\cdot)$: local logit first stage



Note: Boxplots of the RMSE values of the nonparametric estimator of $m(\cdot)$ in 1,000 replications.

4.4 Revisiting Knowledge Capital Production Function Estimation

This section presents the empirical application of the proposed estimation strategy to Italian manufacturing data. The data and the instrument-selection procedure are described in Appendices C.1 and C.2, respectively. The focus here is on the comparison between parametric and semiparametric estimates.

We begin with the parametric estimation of the production function proposed in (4.1), examining the effects of the conventional inputs, labor ℓ and capital k , together with process innovation, IP , on output. Five alternative specifications are considered. Column (i) reports the standard OLS estimates, treating all regressors as exogenous. Columns (ii) and (iii) address the endogeneity of the binary regressor IP by employing, respectively, the semiparametric single-index estimator (SP-IV) and the nonparametric local likelihood logit estimator (NP-IV) to obtain the fitted probabilities used as generated instruments in the second stage. Columns (iv) and (v) extend this framework by additionally treating labor as endogenous, using the same generated instruments for IP together with a separate set of instruments for ℓ .

The results are broadly consistent across specifications. The labor elasticity remains remarkably stable, ranging from 0.729 to 0.768 across all IV estimates, while the capital elasticity declines modestly, from 0.185 under OLS to approximately 0.159–0.163 when both sources of endogeneity are taken into account. In all cases, the estimated elasticity of scale lies slightly below unity, pointing to mild decreasing returns to scale. The most striking pattern concerns the coefficient on IP . Under OLS, its estimate is small and statistically indistinguishable from zero (0.020); once endogeneity is explicitly addressed, however, the estimated effect increases substantially—to 0.293 under SP-IV and 0.209 under NP-

IV when only IP is treated as endogenous (columns ii–iii), and to 0.241 and 0.170, respectively, when both IP and labor are treated as endogenous (columns iv–v). These findings are broadly consistent with the pattern documented in the previous empirical literature based on parametric specifications and CIS data. Parisi et al. (2006), adopting the same Cobb-Douglas framework with an endogenous innovation dummy and estimating the model by 2SLS on Italian manufacturing data, report an innovation coefficient of 0.16, which is broadly in line with our NP-IV estimates under joint endogeneity. Chudnovsky et al. (2006), using a panel data approach, obtain estimates of comparable magnitude. At the upper end of the range, Mairesse & Robin (2008) instrument for innovation by using predicted probabilities from the innovation equation and reports coefficients between 0.35 for manufacturing and 0.59 for services. Our SP-IV estimates, ranging from 0.241 to 0.293, lie comfortably within the lower portion of this interval for the manufacturing sector.

Table 4.3: Parametric structural model Estimates

	Exog.	Endog. IP		Endog. ℓ and IP	
	OLS (i)	SP-IV (ii)	NP-IV (iii)	SP-IV (iv)	NP-IV (v)
β_ℓ	0.7500 (0.0165)	0.7293 (0.0195)	0.7345 (0.0187)	0.7676 (0.0636)	0.7670 (0.0635)
β_k	0.1849 (0.0109)	0.1790 (0.0124)	0.1814 (0.0120)	0.1585 (0.0306)	0.1625 (0.0312)
β_{IP}	0.0196 (0.0149)	0.2930 (0.0936)	0.2088 (0.0662)	0.2409 (0.0972)	0.1699 (0.0652)
Intercept	3.4819 (0.1121)	3.4515 (0.1221)	3.4462 (0.1200)	3.2733 (0.1661)	3.2688 (0.1629)
E.S.	0.9349	0.9082	0.9159	0.9261	0.9295
MRTS	4.0562	4.0742	4.0489	4.8428	4.7199
N	2924	2598	2598	1890	1890

Notes: Standard errors are reported in parentheses. E.S. denotes the elasticity of scale, computed as $\hat{\beta}_\ell + \hat{\beta}_k$, while MRTS denotes the marginal rate of technical substitution $\hat{\beta}_\ell/\hat{\beta}_k$. OLS treats the innovation as exogenous. SP-IV relies on the semiparametric Klein–Spady single-index estimator, whereas NP-IV employs the nonparametric first stage estimator of Frölich (2006). The pilot bandwidth is selected according to Silverman’s rule of thumb $\lambda = 1.06 \times n^{-1/4}$, yielding $\lambda = 0.1485$ for $n = 2,598$ (columns ii–iii) and $\lambda = 0.1608$ for $n = 1,890$ (columns iv–v), and is applied uniformly to both the instrument and the sectoral variable.

A direct comparison can also be drawn with Antonioli et al. (2021), who estimate eq. (4.1) using the same Italian manufacturing data. When correcting only for endogeneity of IP , they obtain $\widehat{\beta}_{IP} = 0.291$ with $\widehat{\beta}_{\ell} = 0.730$ and $\widehat{\beta}_k = 0.179$. When both IP and labor are treated as endogenous, the estimated innovation effect declines to $\widehat{\beta}_{IP} = 0.232$. These values are closely aligned with our parametric IV estimates reported in columns (ii)–(iii) and (iv)–(v), respectively. The downward bias observed under OLS is therefore consistent with the broader empirical evidence documenting a systematic underestimation of the returns to innovation when endogeneity is ignored.

Building on the insights obtained from the parametric analysis, we next extend the investigation to the semiparametric partially linear model specified in (4.3).

Table 4.4: Nonparametric structural model estimates

	Exog.	Endog. IP		Endog. ℓ and IP	
	PLM-OLS (i)	PLM-SP-IV (ii)	PLM-NP-IV (iii)	PLM-SP-IV (iv)	PLM-NP-IV (v)
β_{ℓ}	0.6266 (0.0137)	0.6167 (0.0144)	0.6212 (0.0140)	0.5313 (0.0330)	0.5310 (0.0331)
β_k	0.1985 (0.0092)	0.1913 (0.0083)	0.1931 (0.0078)	0.2213 (0.0316)	0.2208 (0.0317)
β_{IP}	0.0072 (0.0186)	0.4036 (0.1591)	0.3373 (0.1145)	0.2846 (0.1487)	0.2896 (0.1250)
E.S.	0.8251	0.8080	0.8143	0.7798	0.7793
MRTS	3.1565	3.2233	3.2176	2.5247	2.5295
N	2924	2598	2598	1890	1890

Notes: Standard errors are reported in parentheses. E.S. denotes the elasticity of scale evaluated at the sample mean, while MRTS denotes the marginal rate of technical substitution, $\widehat{\beta}_{\ell}/\widehat{\beta}_k$, also evaluated at the mean. PLM-OLS treats both innovation and labor as exogenous. Columns (ii)–(iii) instrument only for IP , using fitted probabilities obtained from the Klein–Spady semiparametric single-index estimator and the nonparametric first stage estimator of Frölich (2006), respectively. Columns (iv)–(v) additionally address the endogeneity of labor via a CF approach, under the same two first-stage estimators. For the Frölich (2006) first stage, the pilot bandwidth is selected according to Silverman’s rule of thumb $\lambda = 1.06 \times n^{-1/4}$, yielding $\lambda = 0.1485$ for $n = 2,598$ (columns ii–iii) and $\lambda = 0.1608$ for $n = 1,890$ (columns iv–v), and is applied to both the instrument and the sectoral variable. Reported input elasticities correspond to sample means of the pointwise estimates.

Table 4.4 reports estimates of the semiparametric partially linear model under three identification strategies. Column (i) treats both IP and labor as exogenous (PLM-OLS). Columns (ii) and (iii) address the endogeneity of IP alone, using fitted probabilities obtained from the single-index and the local likelihood logit

estimator, respectively. Columns (iv) and (v) extend the analysis to the joint endogeneity of both IP and labor. The OLS estimate of β_{IP} is close to zero and statistically insignificant. Correcting only for the endogeneity of IP increases the estimated effect substantially, to approximately 0.40–0.34 (columns ii–iii), and renders it statistically significant at conventional levels. When labor endogeneity is additionally addressed, the estimate moderates to around 0.28–0.29 (columns iv–v), remaining both significant and stable across semiparametric and nonparametric first-stage specifications.

The behavior of the input elasticities is also informative. The labor elasticity declines monotonically as the endogeneity correction is progressively extended: from $\widehat{\beta}_\ell \approx 0.63$ under OLS, to approximately 0.62 when only IP endogeneity is addressed, and further to about 0.53 under joint endogeneity. By contrast, the capital elasticity exhibits the opposite pattern, increasing from roughly 0.20 to 0.22. As a consequence, the elasticity of scale falls to about 0.78 under full endogeneity correction, indicating decreasing returns to scale, while the marginal rate of technical substitution declines from roughly 3.16 – 3.22 to approximately 2.52. A brief discussion of the sector-specific productivity differentials implied by the estimated intercept shifts $\widehat{\beta}_j$ is provided in Appendix C.3.

Placing these results in the context of the existing literature, the PLM estimates of the innovation effect are systematically larger than their parametric counterparts obtained here and in related studies. Under joint endogeneity, the PLM yields $\widehat{\beta}_{IP} \approx 0.28\text{--}0.29$, compared to 0.170–0.241 in the parametric IV estimates reported in columns (iv)–(v). When only IP endogeneity is addressed, the PLM estimates (0.34–0.40) exceed the upper bound of the manufacturing range reported by Mairesse & Robin (2008) (0.35) and are substantially larger than the benchmark of 0.16 for Italian manufacturing reported by Parisi et al. (2006). Notably, the magnitude of the estimated effect declines from approximately 0.37 to

0.29 once labor endogeneity is also accounted for.

The broader empirical literature on process innovation provides useful context for interpreting these magnitudes. Estimates of the effect of *IP* are known to vary considerably across studies and estimation strategies. For instance, Duguet (2006) report estimates ranging from 0.13 to 0.30, with statistical significance depending sensitively on the specification. Griffith et al. (2006), analyzing data for four European countries, generally find no significant effect of process innovation, whereas Parisi et al. (2006) and Chudnovsky et al. (2006) identify it as the only innovation variable with a significant impact. Such variability has been widely attributed to differences in modelling assumptions and estimation techniques.

A direct comparison with Antonioli et al. (2021) further underscores this point. Using the same Italian manufacturing dataset, their additive specification yields $\widehat{\beta}_{IP} = 0.291$ when correcting only for the endogeneity of *IP*, closely matching our parametric SP-IV estimate in column (ii). When both *IP* and labor are treated as endogenous, their estimate declines to $\widehat{\beta}_{IP} = 0.232$, again consistent with our parametric range of 0.170–0.241 in columns (iv)–(v). By contrast, the PLM estimates under the corresponding identification strategies are moderately higher—0.34–0.40 when only *IP* is instrumented, and 0.28–0.29 under joint endogeneity—highlighting the role of functional form flexibility in capturing the contribution of innovation to firm-level productivity.

Table 4.5: PLM-NP-IV estimates: sensitivity to sector bandwidth

	$h = 0.1$	$h = 0.2$	$h = 0.3$	$h = 0.4$	$h = 0.5$	$h = 0.6$	$h = 0.7$	$h = 0.8$	$h = 0.9$
β_k	0.1993 (0.0091)	0.1982 (0.0093)	0.1974 (0.0095)	0.1969 (0.0096)	0.1964 (0.0097)	0.1961 (0.0098)	0.1958 (0.0098)	0.1956 (0.0099)	0.1954 (0.0099)
β_ℓ	0.5614 (0.0134)	0.5600 (0.0136)	0.5591 (0.0138)	0.5584 (0.0139)	0.5579 (0.0140)	0.5575 (0.0141)	0.5572 (0.0142)	0.5569 (0.0143)	0.5567 (0.0143)
E.S.	0.7606	0.7582	0.7565	0.7553	0.7544	0.7536	0.7530	0.7525	0.7521
$\widehat{\beta}_{IP}$	0.2625 (0.1137)	0.3036 (0.1312)	0.3309 (0.1438)	0.3505 (0.1535)	0.3652 (0.1611)	0.3768 (0.1672)	0.3861 (0.1722)	0.3937 (0.1764)	0.4000 (0.1799)
N	2598								

Notes: Standard errors are reported in parentheses. Each column presents PLM-IV-NP estimates based on the nonparametric first stage estimator of Frölich (2006), varying the bandwidth h applied to the sectoral discrete variable while holding the instrument bandwidth fixed at $h_{IV} = 0.1608$. The reported input elasticities, $\widehat{\beta}_k$ and $\widehat{\beta}_\ell$, correspond to sample means of the pointwise estimates. E.S. denotes the mean elasticity of scale.

Table 4.5 investigates the sensitivity of the PLM-NP-IV estimates to the choice of bandwidth λ applied to the sectoral discrete variable in the local likelihood logit first stage, while keeping the instrument bandwidth fixed at $\lambda_{IV} = 0.1608$. The estimated innovation effect, $\widehat{\beta}_{IP}$, increases monotonically from 0.26 at $\lambda = 0.1$ to 0.40 at $\lambda = 0.9$. This pattern reflects progressive pooling across sectors induced by larger bandwidths, which smooths the estimated first-stage conditional probabilities and, consequently, alters the resulting IV projection. By contrast, the input elasticities, $\widehat{\beta}_k$ and $\widehat{\beta}_\ell$, remain remarkably stable across bandwidth choices, with only negligible variation at the third decimal place, indicating that the estimated production technology is largely robust to the degree of smoothing applied in the first stage.

Table 4.6: Estimates of input's elasticity: mean and quantiles

Method	Capital					Labor				
	$\widehat{\beta}_k$	$\widehat{\beta}_k(\bar{x})$	k_{25}	k_{50}	k_{75}	$\widehat{\beta}_\ell$	$\widehat{\beta}_\ell(\bar{x})$	ℓ_{25}	ℓ_{50}	ℓ_{75}
PLM-OLS	0.1985	0.2283	0.1347	0.1952	0.2497	0.6266	0.5956	0.5264	0.6647	0.7507
PLM-IV-SP	0.1948	0.2213	0.1372	0.1999	0.2459	0.5740	0.5584	0.5057	0.5924	0.6688
PLM-IV-NP	0.1946	0.2208	0.1372	0.2000	0.2457	0.5738	0.5585	0.5052	0.5921	0.6679

Notes: $\widehat{\beta}$ denotes the sample mean of pointwise elasticity estimates, while $\widehat{\beta}(\bar{x})$ denotes the elasticity evaluated at the sample mean of the input distribution. Reported quantiles refer to the empirical distribution of the pointwise elasticity estimates.

Table 4.6 reports the distribution of firm-level input elasticities across the three estimators. Under PLM-OLS, both capital and labor elasticities exhibit moderate cross-sectional dispersion, with interquartile ranges of approximately 0.11 for $\widehat{\beta}_k$

and 0.22 for $\widehat{\beta}_\ell$. The IV specifications display a broadly similar degree of heterogeneity. In particular, the interquartile range for $\hat{\beta}_k$ remains close to 0.11, while that for $\widehat{\beta}_\ell$ is approximately 0.16. These results indicate that the nonparametric component $m(k, l)$ captures substantial firm-level variation in input elasticities, even after correcting for the endogeneity of innovation.

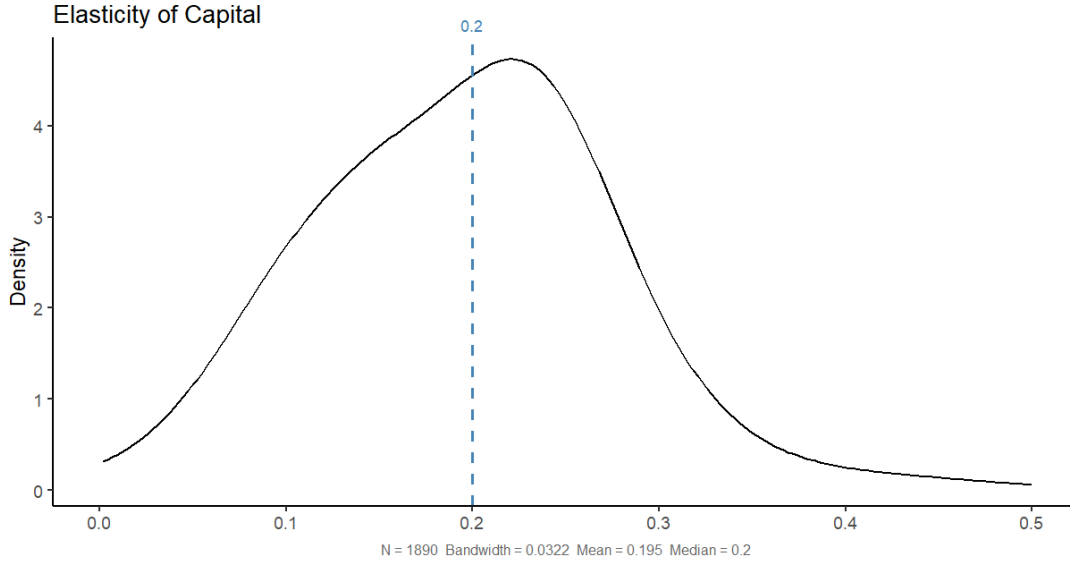
Table 4.7: Elasticities of scale: mean and quantiles

Method	$\overline{\text{ES}}$	$\widehat{\text{ES}}(\bar{x})$	ES_{25}	ES_{50}	ES_{75}
PLM-OLS	0.8251	0.8240	0.7036	0.8553	0.9615
PLM-IV-SP	0.7688	0.7798	0.6751	0.7896	0.9006
PLM-IV-NP	0.7684	0.7793	0.6746	0.7887	0.9001

Notes: $\overline{\text{ES}}$ denotes the sample mean of the pointwise elasticity of scale estimates, while $\widehat{\text{ES}}(\bar{x})$ denotes the elasticity of scale evaluated at the sample mean of the input distribution.

Table 4.7 summarises the distribution of the firm-level elasticities of scale. Across all three estimators, the mean elasticity of scale is below unity, indicating decreasing returns to scale at the sample mean. The PLM-OLS specification exhibits a relatively wide distribution, with an interquartile range of approximately 0.26, spanning from 0.70 at the 25th percentile to 0.96 at the 75th percentile. By contrast, the IV specifications display a more concentrated distribution. Both PLM-IV-SP and PLM-IV-NP yield an interquartile range of roughly 0.23, with values ranging from about 0.68 at the lower quartile to 0.90 at the upper quartile, indicating a more compressed cross-sectional dispersion once endogeneity is accounted for.

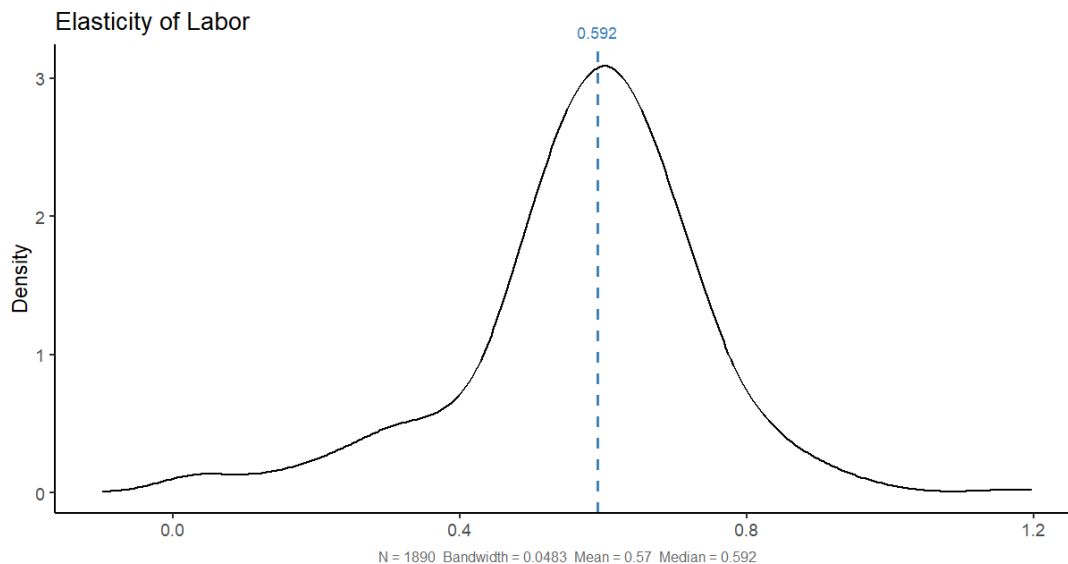
Figure 4.3: Empirical distribution of the capital elasticity: PLM-NP-IV (v) estimates



Note: The figure presents the kernel density estimate of the firm-level capital elasticity $\hat{\beta}_k$, obtained from the semiparametric partially linear model with mixed endogeneity, where both process innovation and capital are treated as endogenous. The dashed vertical line indicates the median, as reported on the horizontal axis. The sample mean is $\bar{\beta}_k = 0.195$. The estimation is based on $N = 1,890$ observations, and the bandwidth is selected according to Silverman's rule of thumb ($h = 0.032$).

Figure 4.3 depicts the kernel density of the firm-level capital elasticity, $\hat{\beta}_k$, estimated under the mixed-endogeneity PLM-IV specification ($N = 1,890$). The distribution is approximately unimodal and exhibits right skewness, with a median of 0.20 and a sample mean of $\bar{\beta}_k = 0.195$. Most of the mass is concentrated within the interval $[0.10, 0.30]$, while the presence of a pronounced right tail points to a subset of firms characterized by relatively more capital-intensive production technologies.

Figure 4.4: Empirical distribution of the labor elasticity: PLM-NP-IV (v) estimates



Note: The figure presents the kernel density estimate of the firm-level labor elasticity, $\hat{\beta}_l$, obtained from the semiparametric partially linear model with mixed endogeneity, where both process innovation and capital are treated as endogenous. The dashed vertical line indicates the median, as reported on the horizontal axis. The sample mean is $\bar{\beta}_l = 0.570$. The estimation is based on $N = 1,890$ observations, and the bandwidth is selected according to Silverman's rule of thumb ($h = 0.048$).

Figure 4.4 displays the kernel density of the firm-level labor elasticity, $\hat{\beta}_\ell$, obtained from the same mixed-endogeneity specification. The distribution is approximately unimodal and exhibit left skewness, with a median of 0.592 and a sample mean of $\bar{\beta}_\ell = 0.570$. The bulk of the distribution lies between 0.40 and 0.80, with a relatively long left tail indicating a non-negligible share of firms for which labor contributes less than half of marginal output. Taken together, Figures 4.3 and 4.4 provide clear evidence that the nonparametric component of the PLM captures meaningful firm-level heterogeneity in the production technology that would be obscured under a fully parametric specification imposing constant input elasticities across firms.

4.5 Conclusions

We propose a novel estimator for partially linear models in which both a continuous and a binary endogenous regressor enter simultaneously. The framework extends the standard dummy endogenous variable model by allowing both endogenous and exogenous continuous covariates to enter through an unknown smooth function, while preserving a constant structural shift associated to the dummy endogenous variable. The conditional probability $p_i = P(W_i^d = 1 \mid Z_i)$, estimated using alternative semi- and nonparametric first-stage methods, is employed as a generated instrument for W_i^d , whereas nonparametric residuals from the reduced-form equation for the continuous endogenous regressor are used to control for the endogeneity of W_i^c .

The Monte Carlo results demonstrate that the proposed estimator exhibits satisfactory finite-sample performance across a range of endogeneity levels and first-stage specifications. The empirical application to the knowledge capital production function further highlights the practical relevance of the approach. In particular, OLS is shown to substantially underestimate the effect of process innovation on productivity. Once the endogeneity of both process innovation and labor is explicitly addressed within the partially linear framework, the estimated innovation coefficient increases to approximately 0.28–0.29. At the same time, the nonparametric component provides input elasticities to vary across firms, rather than imposing the restrictive assumption of constant elasticities inherent in parametric specifications.

Chapter 5

Conclusions and Future Research

5.1 General Conclusions

The increasing availability of microeconomic data, together with the growing demand for credible causal inference, has led the econometric literature toward progressively more flexible modeling frameworks. This dissertation contributes to this development by proposing a coherent and unified methodological approach to the estimation of partially linear models in the presence of endogeneity. The analysis follows a natural progression, beginning with models featuring continuous endogenous regressors, extending to settings with discrete endogenous variables, and culminating in the mixed case. The central premise underpinning this work is that functional flexibility and causal identification need not be mutually exclusive; rather, they can be jointly achieved through carefully designed sequential procedures that exploit the triangular structure of the system and the availability of valid instruments.

Chapter 1 established the conceptual foundations by providing a critical review of the literature on endogeneity in both parametric and nonparametric contexts. It distinguished among the primary sources of endogeneity (i.e., omitted vari-

ables, measurement error, and simultaneity) and highlighted their implications for identification and estimation. The chapter also examined the trade-off between limited-information and full-information methods, emphasizing the robustness of the former and the efficiency, but also fragility, of the latter. In the nonparametric setting, the recovery of structural relationships was framed as an ill-posed inverse problem, requiring regularization. A key contrast was drawn between nonparametric IV methods, which rely on weaker assumptions but suffer from slower convergence rates, and CF approaches, which impose stronger conditions but attain optimal rates.

Chapter 2 focus on a semiparametric estimator for partially linear models with a continuous endogenous regressor entering the nonparametric component. The proposed estimation procedure proceeds in three stages. First, reduced-form residuals are obtained via local linear regression using mixed product kernels capable of handling both continuous and discrete instruments. Second, The Robinson's transformation is employed to eliminate the nonparametric component through orthogonalization, yielding a $\sqrt{N}$ -consistent and asymptotically normal estimator for the parametric coefficients. Third, the structural function is recovered via marginal integration, implemented efficiently using a Monte Carlo approximation. Simulation results confirm that the estimator performs well in finite samples, with bias, variance, and RMSE declining steadily as sample size increases, even under substantial endogeneity.

Chapter 3 turned to the more challenging case of a dummy endogenous regressor entering the parametric component. Standard CF methods are not directly applicable in this context due to the discrete nature of the variable, the unobservability of the reduced-form error, and the pitfalls of forbidden regression with using fitted values from nonlinear first-stage models. To address these issues, the chapter proposed an alternative strategy based on using the conditional probability

$p_i = P(W_i^d = 1 \mid Z_i)$, estimated using alternative parametric (probit and logit), semiparametric (Klein–Spady), and fully nonparametric (local likelihood logit) methods. Monte Carlo results show that parametric first-stage estimators exhibit substantial dispersion and instability in small samples, whereas semiparametric and nonparametric estimators deliver more stable performance, with low bias and steadily decreasing variance as the sample size increases. The empirical application to the knowledge production function confirms the relevance of the approach: treating process innovation as exogenous yields a small and statistically insignificant estimate ($\hat{\gamma} \approx 0.020$), while addressing endogeneity increases the estimated effect to approximately 0.26 – 0.30 under parametric IV specifications and to 0.40 – 0.42 under semiparametric and nonparametric approaches. The associated elasticity of scale declines from approximately 0.91 – 0.93 in the parametric specification to around 0.81 in the partially linear model, reflecting the impact of relaxing functional form restrictions.

Chapter 4 unified the two preceding strands by considering a partially linear model in which a continuous endogenous variable enters the nonparametric component alongside the exogenous regressors, while the parametric component includes both exogenous regressors and a dummy endogenous variable. The proposed strategy treats continuous endogeneity via a CF—augmenting the model with nonparametric reduced-form residuals—and discrete endogeneity via IV estimation using generated probabilities. The resulting three-stage procedure preserves $\sqrt{N}$ -consistency for the parametric coefficients and the optimal nonparametric convergence rate for $m(\cdot)$. The empirical application on a sample of 1,890 manufacturing firms yields a coherent pattern: correcting for the endogeneity of process innovation alone raises the estimated effect from near zero to approximately 0.34 – 0.40, while extending the correction to the joint endogeneity of labor moderates the innovation coefficient to approximately 0.28 – 0.29.

At the same time, the labor elasticity declines from approximately 0.75 to 0.53, while the capital elasticity increases from about 0.20 to 0.22, leading to a reduction in the elasticity of scale to approximately 0.78. The nonparametric component further reveals substantial heterogeneity in production technology, with interquartile ranges of 0.11 for capital and 0.16 for labour elasticities, underscoring the importance of allowing for firm-level variation that parametric specifications cannot capture.

Taken together, the dissertation makes three main contributions. First, it extends the CF approach to semiparametric models with mixed data types. Second, it introduces a novel estimator for discrete endogenous regressors that avoids both the forbidden regression problem and reliance on strong distributional assumptions. Third, it provides a unified framework that accommodates simultaneous continuous and discrete endogeneity within partially linear models.

5.2 Future Research

A future line of investigation concerns the analysis of localized technical change (Atkinson & Stiglitz 1969) through a varying coefficient instrumental variable estimator. This was indeed the original aim of the present dissertation. The empirical application in Chapters 3 and 4 incorporates a binary indicator of process innovation to account for the effect of technical change as a structural shift, consistent with the *Hicksian* notion of neutral technical change (Hicks 1932). To depart from this restriction, an alternative would be to consider a varying coefficient model within an IV estimation strategy, allowing the effect of technical change to be non-neutral, while jointly addressing the endogeneity of both continuous and discrete variables entering the nonparametric component.

Appendix A

Chapter 2

A.1 Asymptotic Properties - Semiparametric Models with Continuous Endogenous Variables

In this subsection, we provide the asymptotic behavior of the estimators proposed in equations (2.21) and (2.23), which correspond to the parametric component β and the nonparametric component $m(\cdot)$, respectively.

To establish the asymptotic properties of the estimator for β in partially linear models, we need to ensure that the nonparametric part of the regression model does not distort the behavior of the parametric estimator since its estimation procedure relies on projecting out the nonparametric component. For this projection to be well-behaved, we follow Robinson (1988) and impose the following definition that enables us to ensure that the nuisance function exhibits sufficient smoothness for accurate nonparametric estimation and satisfies moment conditions that guarantee the convergence of empirical processes. Therefore, this definition ensures that the resulting parametric estimator yields reliable inference, especially root- N consistency and asymptotic normality.

Definition A.1 $\mathcal{G}_\mu^\alpha$, $\alpha > 0$, $\mu > 0$, is the class of functions $g : \mathcal{R}^d \rightarrow \mathcal{R}$ satisfying the following conditions:

- g is $(\rho - 1)$ -times partially differentiable, for $\rho - 1 \leq \mu < \rho$;
- for some $\rho_0 > 0$, $\sup_{y \in \phi_{z\rho_0}} |g(y) - g(z) - Q_g(y, z)|/|y - z|^\mu \leq G_g(z)$ for all z , where $\phi_{z\rho_0} = \{y : |y - z| < \rho_0\}$;
- $Q_g = 0$ when $\rho = 1$; Q_g is a $(\rho - 1)$ th degree homogenous polynomial in $y - z$ with coefficients the partial derivatives of g at z of orders 1 through $\rho - 1$ when $\rho > 1$;
- $g(z)$, its partial derivatives of order $\rho - 1$ and less, and $G_g(z)$ have finite α th moments.

Under Definition A.1, the functions belonging to the class $\mathcal{G}_\mu^\alpha$ are locally approximable by a Taylor expansion of order $\rho - 1$, where ρ is the smallest integer strictly greater than μ . Also, such functions must be continuously differentiable up to order $\rho - 1$ and the approximation error is required to decay at a rate governed by $\mu - \rho + 1$, thereby ensuring a form of local Hölder continuity. Additionally, the function itself, its derivatives, and the bound on the approximation error must all have finite α th moments. To sum up, this framework generalizes classical smoothness conditions by combining local regularity with global integrability, allowing for unbounded or only weakly smooth functions.

We also define $\xi_{2i} \equiv \xi_2(X_{1i}, W_i^c, u_i) = E[X_{2i}|X_{1i}, W_i^c, u_i]$, $\eta_i = X_{2i} - \xi_{2i}$, $\kappa = \int k(v)^2 dv$ and $\kappa_2 = \int k(v)v^2 dv$. Also, $\|\cdot\|$ denotes the Euclidean norm and the following conditions will be used to establish the $\sqrt{N}$ -consistency of $\widehat{\beta}$.

Assumption A.1

- (i) $\{Y_i, X_{1i}, X_{2i}, W_i^c, \epsilon_i, u_i\}$ are independent and identically distributed (i.i.d.). Also, $\epsilon_i = Y_i - m_0(X_{1i}, W_i^c, u_i) - \beta^\top X_{2i}$ is the error term with zero mean and $E(|\epsilon_i|^{4(1+\delta)}) < \infty$, $E(\|\eta_i\|^{4(1+\delta)}) < \infty$, and $E(|u_i|^{4(1+\delta)}) < \infty$, for some $\delta > 0$.

- (ii) For some $\delta > 0$, X_{1i} and W_i^c has finite $4(1+\delta)$ th, and Z_t has finite $8(1+\delta)$ th moments, $f \in \mathcal{G}_{v-1}^\infty$, where $f(\cdot)$ is the joint density function of (X_{1i}, W_i^c, u_i) , $m(\cdot) \in \mathcal{G}_v^4$, and $\xi_2(\cdot) \in \mathcal{G}_v^4$, where $v \geq 2$ is an integer. Furthermore, all smooth functions f , m , and ξ_2 satisfy some (global) Lipschitz-type condition: $|g(y) - g(z)| \leq D_g(z)\|y - z\|$ for all y and z , where $D_g(z)$ has finite fourth moments and $g \in \{f, m, \text{ or } \xi_2\}$.
- (iii) For c_n being sequences that are either bounded or diverging slowly to infinity, $\sup_{|z| \leq c_n} |\widehat{g}(z) - g(z)| = O_p\left((\log N / N h_1^q)^{1/2} + h_1^2\right)$ when Z_{1i} is a continuous variable and $\sup_{|z| \leq c_n} |\widehat{g}(z) - g(z)| = O_p\left((\log N / N h_1^{q_0})^{1/2} + h_1^2\right)$ when Z_{1i} is discrete and $q_0 = q - q_{z_1}$.
- (iv) $K(\cdot)$ is a product kernel and the univariate kernel $k(\cdot)$ is a nonnegative and bounded function satisfying $\int k(u) du = 1$, $\int k(u) u du = 0$, and $\int k(u) u^2 du > 0$ and finite.
- (v) $k(\cdot)$ has second-order continuous derivatives; $k(\cdot)$ and its derivatives (up to second order) are all bounded. Moreover, $k(\cdot)$ satisfies a (global) Lipschitz-type condition: $|k(u+v) - k(u) - k^{(1)}(u)v| \leq \frac{1}{2}H(u)v^2$ for all $u, v \in \mathcal{R}$, where $k^{(1)}(u) = \partial k(u) / \partial u$ and $H(\cdot)$ is a bounded function. To simplify the notation, we also assume $k(u) \leq H(u)$ and $k^{(1)}(u) \leq H(u)$.
- (vi) As $N \rightarrow \infty$, $h_2 \rightarrow 0$, $N h_2^{2(q_{x_1} + q_{z_1} + 1)} \rightarrow \infty$ and $N h_2^{4v} \rightarrow 0$ for some $v \geq 2$.
- (vii) The matrix function $E[(X_{2i} - \xi_{2i})(X_{2i} - \xi_{2i})^\top f^2(X_{1i}, W_{1i}^c, u_i)]$ is non-singular.

These assumptions are fairly standard in the semiparametric literature, but some clarifications are warranted. Assumptions A.1(i) and (ii) are some mild smoothness and moments conditions. Also, Assumption A.1(ii) is stronger than the corresponding condition in Robinson (1988). Specifically, it is assumed that the density function is bounded and at least first-order partially differentiable with

a Lipschitz-continuous remainder. This stronger regularity condition is required for the estimation of the nonlinear component, though it is not necessary for establishing the asymptotic properties of $\widehat{\beta}$. This assumption is only essential for the consistency and asymptotic behavior of the third-stage estimator, $\widehat{m}(\cdot)$. Assumption A.1(iii) assumes that $\widehat{g}(\cdot)$ is a consistent estimator of $g(\cdot)$. Assumption A.1(iv) is a standard condition in the nonparametric literature where is assumed that we work with a nonnegative second-order kernel function, whereas Assumption A.1(v) imposes some smoothness conditions on the kernel functions that is quite common when generated regressors are considered in $k(\cdot)$ (see Li & Wooldridge (2002) for further details). Assumption A.1(vi) implies that $(2\nu > q_{x_1} + q_{z_1} + 1)$ or $(2\nu > q_{x_1} + q_{z_1} + 2)$ (because ν is an integer), which is equivalent to the condition $(\nu > (q_{x_1} + q_{z_1} + 1)/2)$, as in Robinson (1988) or Li & Stengos (1996). Consequently, a standard second-order kernel ($\nu = 2$) may be used provided that $(q_{x_1} + q_{z_1} + 1) \leq 3$. Alternatively, Li (1996) proposes a weaker condition under which the use of a standard second-order kernel remains valid. Nonetheless, in this chapter we adopt the stronger condition stated in Assumption A.1(vi), as it facilitates a relatively straightforward proof of the asymptotic properties of the proposed estimators.

The following theorem collects the main asymptotic features of the parametric estimator $\widehat{\beta}$ proposed in the second stage.

Theorem 1.1 *Let $\eta_i = X_{2i} - \xi_{2i}$ and $\xi_{2i} = E(X_{2i}|X_{1i}, W_i^c, u_i)$. Under Assumption A.1, as $N \rightarrow \infty$ we get*

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Psi_a^{-1} \Omega_a \Psi_a^{-1}),$$

provided Ψ_a is positive definite, where $\Psi_a = E(\eta_i \eta_i^\top f_i^2)$, $\Omega_a = \sigma_\epsilon^2 E(\eta_i \eta_i^\top f_i^4)$, and $f_i = f(X_{1i}, W_i^c, u_i)$.

Theorem 1.1 demonstrates that the estimator $\widehat{\beta}$ is asymptotically normal and $\sqrt{N}$ -

consistent, with its limiting distribution characterized by a sandwich-form covariance matrix. Given the previously described two-stage estimation approach, this result formalizes the impact of using residuals in the second stage to control for the endogeneity issue. Specifically, the term $\eta_i = X_{2i} - \xi_{2i}$ ensures that the variation used to identify β is orthogonal to the nonparametric component. This orthogonalization is crucial for achieving unbiased estimation in the presence of potentially endogenous or high-dimensional controls. The form of the asymptotic variance reflects the influence of the first-stage estimation on the overall efficiency of $\widehat{\beta}$. The result in Theorem 1.1 builds upon the theoretical frameworks developed for Theorem 2.1 in Li & Wooldridge (2002) for a different setting, ensuring that standard inference procedures remain valid. The detailed proof is available upon request.

Considering now the asymptotic properties of the nonparametric estimator $\widehat{m}(x_1, w_1^c)$. The following regularity conditions are imposed to ensure its desirable asymptotic properties.

Assumption A.2

- (i) $\omega(\cdot)$ is a positive weighting function that is defined on the compact support $\mathbb{W}$, twice continuously differentiable, $\int_{\mathbb{W}} \omega(u) du = 1$, and $\int_{\mathbb{W}} m(u) \omega(u) du = 0$.
- (ii) $k(\cdot)$ is a second-order kernel that is positive, bounded, symmetric, and defined on the support $\mathbb{W}$.
- (iii) $h_3 = h_0 N^{-1/(4+q_{x_1}+q_{z_1})}$, where h_0 is a positive bandwidth.

Note that the positiveness of the weighting function $\omega(\cdot)$ is not essential for the asymptotic analysis and it is assumed here to facilitate the use of Monte Carlo methods for numerical integration. The remaining conditions on $\omega(\cdot)$ primarily serve normalization purposes. In particular, if $\int_{\mathbb{W}} \omega(u) du \neq 1$, then the estimator

$\widehat{m}(x_1, w^c)$ is identified only up to a multiplicative constant. Similarly, if the condition $\int_{\mathbb{W}} m(u)\omega(u)du = 0$ is not imposed, the estimator $\widehat{m}(\cdot)$ may be identified only up to an additive constant (see Qian & Wang (2012) for a deeper explanation). Let $\widehat{m}(x_1, w^c)$ be defined as in (2.24), the following theorem contains the main asymptotic properties of the nonparametric estimator.

Theorem 1.2 *Under Assumptions A.1 and A.2, as $N \rightarrow \infty$ we get*

$$N^{2/(4+q_{x_1}+q_{z_1})}\{\widehat{m}(x_1, w^c) - m(x_1, w^c)\} \xrightarrow{d} \mathcal{N}(B_a(x_1, w^c), V_a(x_1, w^c))$$

where

$$B_a(x_1, w^c) = \frac{h_0^2 \kappa_2^{(q_{x_1}+q_{z_1})}}{2} \left[\text{tr}\{\mathcal{H}_m(x_1, w^c)\} - \int_{\mathbb{W}} \text{tr}\{\mathcal{H}_m(u)\}\omega(u)du \right],$$

$$V_a(x_1, w^c) = \frac{\kappa^{(q_{x_1}+q_{z_1})}\sigma_\epsilon^2}{h_0^{q_{x_1}+q_{z_1}}} \left(\int_{\mathbb{W}} \frac{\omega^2(u)}{f(x_1, w^c, u)} du \right),$$

and $\mathcal{H}_m(x_1, w^c)$ is the Hessian matrix of $m(x_1, w^c)$.

Theorem 1.2 characterizes the asymptotic distribution of the nonparametric estimator $\widehat{m}(x_1, w^c)$, showing that it converges at rate $N^{2/(4+q_{x_1}+q_{z_1})}$ to a Normal distribution with a non-zero bias term $B_a(x_1, w^c)$ and variance $V_a(x_1, w^c)$. This result, aligned with the marginal local linear estimation framework proposed by Qian & Wang (2012), highlights the classical bias–variance trade-off in kernel-based estimation. The bias term depends on the curvature of the unknown regression function $m(\cdot)$, as captured by the trace of its Hessian matrix, adjusted to ensure identification through integration against the weighting function $\omega(\cdot)$. The asymptotic variance reflects both the smoothing properties of the kernel and the conditional error variance σ_ϵ^2 , scaled by the local density and the weighting function. These results confirm that the estimator is pointwise asymptotically normal but exhibits local bias. The proof of this theorem is straightforward combining the proof schemes of Theorem 2 in Qian & Wang (2012) and in Lemma

B.1 in Su & Ullah (2008). The detailed proof is available upon request.

Appendix B

Chapter 3

B.1 Estimation Procedure - Alternative First-Stage Estimators for Fitted Probabilities

In our simulations, we compare alternative estimators of $P(W_i^d = 1 \mid Z_i = z)$ that span from fully parametric to fully nonparametric approaches. At the parametric end, we use standard probit and logit models. At an intermediate level, we employ the semiparametric single-index estimator of Klein & Spady (1993), which relaxes the specification of the link function but still restricts covariates to enter through a linear index. Across nonparametric methods, direct smoothers such as Nadaraya-Watson or local linear regression approximate probabilities by local averages of binary outcomes. While natural, this approach is often unstable: it can yield predictions outside the $[0, 1]$ interval and behaves poorly near boundaries. Moreover, these methods do not account for the heteroskedasticity structure of binary outcomes, leading to inefficient and potentially biased estimators, particularly near the boundaries of the support of Z_i where standard nonparametric estimators tend to perform poorly due to boundary bias (see (Li & J. S. Racine 2007; Hastie & Tibshirani 1990)). For these reasons, we consider

the local-likelihood logit estimator of Frölich (2006), which allows the relationship between covariates and probabilities to be fully flexible while maintaining key advantages of parametric logit models. By imposing a local likelihood with a logit link, it ensures that predicted probabilities remain within the unit interval and accounts for the intrinsic heteroskedasticity of binary outcomes, where $\text{Var}(W_i^d | Z_i = z) = g(z)(1 - g(z))$. This sets it apart from both parametric models, which impose global functional form restrictions, and semiparametric single-index approaches, which still require covariates to enter linearly.

In the end, we consider four alternative estimators for $p_i = P(W_i^d = 1 | Z_{2i})$. First, we estimate p_i using parametric binary response models, namely probit and logit (i.e., $p_i = \Lambda(Z_{2i}^\top \pi)$, where $\Lambda(\cdot)$ is the corresponding CDF). Second, we estimate the fitted probabilities using semiparametric/nonparametric methods (Klein & Spady 1993; Frölich 2006; P. Hall, J. Racine, et al. 2004).

Parametric estimators. In the probit model, the latent index is linear, $\theta(z) = Z_i^\top \delta$, and the link function is the standard normal CDF:

$$\Psi(z, \theta_z) = \Phi(Z_i^\top \delta), \quad (\text{B.1})$$

with δ obtained by maximum likelihood (ML). The fitted probabilities are

$$\widehat{p}_i = \Phi(Z_i^\top \widehat{\delta}). \quad (\text{B.2})$$

In the logit model, the logistic link is used:

$$\Psi(z, \theta_z) = \Lambda(Z_i^\top \delta) = \frac{\exp(Z_i^\top \delta)}{1 + \exp(Z_i^\top \delta)}, \quad (\text{B.3})$$

with δ obtained by maximum likelihood (ML). The fitted probabilities are

$$\widehat{p}_i = \frac{\exp(Z_i^\top \widehat{\delta})}{1 + \exp(Z_i^\top \widehat{\delta})}. \quad (\text{B.4})$$

In both models, the latent index is fully determined by the parameter vector δ rather than allowed to vary freely with z .

Semiparametric single-index estimator. To relax the parametric specification of the link function, we use the Klein & Spady (1993) estimator, which maintains a linear index $Z_i^\top \delta$ but treats the link function $G(\cdot)$ as unknown and estimates it nonparametrically. The model is $W_i^d = G(Z_i^\top \delta_0) + \varepsilon_i$, where $G(\cdot)$ is an unknown link function and ε_i satisfies $E[\varepsilon_i | Z_i] = 0$. For a candidate parameter vector δ and bandwidth h , the method constructs the index values $v_i = Z_i^\top \delta$ for each observation. The conditional probability $P(W_i^d = 1 | v_i)$ is then estimated by a leave-one-out local constant (Nadaraya–Watson) kernel estimator:

$$\widehat{P}_{-i}(\delta, h) = \frac{\sum_{j \neq i} K\left(\frac{v_i - v_j}{h}\right) W_j^d}{\sum_{j \neq i} K\left(\frac{v_i - v_j}{h}\right)}, \quad (\text{B.5})$$

where $K(\cdot)$ is a kernel function (e.g., second-order Gaussian). The estimator $\widehat{\delta}$ and bandwidth $\widehat{h}$ are obtained by jointly maximizing the leave-one-out log-likelihood:

$$(\widehat{\delta}, \widehat{h}) = \arg \max_{\delta, h} \frac{1}{N} \sum_{i=1}^N \left[W_i^d \log \widehat{P}_{-i}(\delta, h) + (1 - W_i^d) \log (1 - \widehat{P}_{-i}(\delta, h)) \right]. \quad (\text{B.6})$$

For identification, the first element of δ is normalized to one ($\delta_1 = 1$), and the regressors must include at least one continuous variable. The fitted probabilities are then

$$\widehat{p}_i = \widehat{P}_{-i}(\widehat{\delta}, \widehat{h}) \quad (\text{B.7})$$

for $i = 1, \dots, N$. This semiparametric approach allows flexible nonlinear effects of the index on the probability while maintaining a parametric linear index for identification and interpretation. Under appropriate regularity conditions, the estimator is consistent and $\sqrt{N}$ -asymptotically normal; see Klein & Spady (1993) for detailed theoretical properties. For the computational implementation, including joint optimization via multistarting and kernel summation routines, see Li & J. S. Racine (2007) and Hayfield & J. S. Racine (2008).

Fully nonparametric estimator. The estimator proposed by Frölich (2006) allows the latent index $\theta(z)$ to vary arbitrarily with z , giving full flexibility:

$$\Psi(z, \theta_z) = G\{\theta(z)\}, \quad (\text{B.8})$$

where $G(\cdot)$ is specified as the logistic link. A critical step is the recovery of the latent index function $\theta(z)$, for which several nonparametric methods have been proposed. One widely cited method is the local nonlinear least squares estimator of Gozalo & Linton (2000), which estimates $\widehat{\theta}_z$ by solving

$$\widehat{\theta}_z = \arg \min_{\theta_z} \sum_{i=1}^N \left[W_i^d - \Psi(Z_i, \theta_z) \right]^2 K_{h_1}(Z_i, z), \quad (\text{B.9})$$

where $K_{h_1}(\cdot)$ is a kernel function with bandwidth h_1 , and $\Psi(Z_i, \theta_z)$ is a smooth predictor of the binary outcome. Although intuitive, this method relies on a squared-error loss, implicitly assumes Gaussian errors, and can be inefficient or inconsistent for Bernoulli outcomes. It may also produce predictions outside the unit interval.

To avoid these shortcomings, we estimate θ_z using a local likelihood method (Staniswalis 1989; Severini & Staniswalis 1994; Fan, N. E. Heckman, et al. 1995; Fan & Gijbels 1996), which extends classical maximum likelihood estimation by incorporating kernel weights to localize the likelihood function. In the binary case, this entails fitting localized Bernoulli likelihoods with link functions such as the logit or probit, ensuring that predicted probabilities remain within $[0, 1]$ while capturing local structure in the data. The concavity of the local likelihood under standard links, such as the logistic, enhances numerical stability and facilitates convergence. Moreover, as the bandwidth increases, the estimator converges to the parametric maximum likelihood estimator, retaining asymptotic efficiency under correct specification while remaining robust in more flexible data-driven settings (see Loader (1996) & Härdle et al. (2000)).

In the local likelihood estimation approach to estimate θ_z Frölich (2006) proposed the following. Let $\{Z_i, W_i^d\}_{i=1}^N$ denote the observed sample. The log-likelihood function is defined as

$$\sum_{i=1}^N \left[W_i^d \log(p(Z_i)) + (1 - W_i^d) \log(1 - p(Z_i)) \right] \quad (\text{B.10})$$

and assuming that the conditional probability of success is modeled as $p(z) = G\{\theta(z)\}$ and the link function is logistic (i.e., $G\{\theta(Z_i)\} = \frac{1}{1+e^{-\theta(Z_i)}}$), the log-likelihood logit function becomes

$$\sum_{i=1}^N \left[W_i^d \log\left(\frac{1}{1+e^{-\theta(Z_i)}}\right) + (1 - W_i^d) \log\left(\frac{1}{1+e^{\theta(Z_i)}}\right) \right]. \quad (\text{B.11})$$

Given the unknown structure of $\theta(Z_i)$, we approximate it locally around a given point z using a Taylor expansion.

Note that an important practical consideration arises when Z_i is a vector of mixed-type instruments (i.e., both continuous or discrete) and the estimation procedure proposed in this section must be adapted to account for the mixed nature of the covariates. Let us denote $Z_i = (Z_i^c, Z_i^d) \in \mathbb{R}^q$, where $Z_i^c \equiv (X_{1i}, X_{2i})$ represents the continuous covariates and forms a vector of dimension $q_x = q_{x_1} + q_{x_2}$, and $Z_i^d \equiv Z_{1i}$ denotes the discrete variables assumed to be of dimension q_{z_1} .

To incorporate this structure into the local estimation framework, we employ a local linear approximation of the unknown function $\theta(Z_i)$, focusing on the continuous component Z_i^c . The approximation around a target point $z = (z^c, z^d)$ is specified as

$$\theta(Z_i) \equiv \theta_0 + (Z_i^c - z^c)^\top \theta_{1,c},$$

where $\theta_0 = \theta(z)$ represents the value of the function at the target point and $\theta_{1,c} = \frac{\partial \theta(z^c)}{\partial z^c}$ is the gradient vector of $\theta(\cdot)$ with respect to the continuous variable $z^c \in \mathbb{R}^{q_x}$. Hence, this formulation allows for local variation in the effect of

the continuous covariates, while treating the discrete component as fixed within each local neighborhood.

Substituting this local linear approximation in (B.11) we propose the following localized log-likelihood function in the presence of both continuous and discrete variables

$$\begin{aligned} \ell_N(\theta_0, \theta_{1,c}) &= \sum_{i=1}^N \left[W_i^d \log \left(\frac{1}{1 + e^{-(\theta_0 + (Z_i^c - z^c)^\top \theta_{1,c})}} \right) + (1 - W_i^d) \log \left(\frac{1}{1 + e^{\theta_0 + (Z_i^c - z^c)^\top \theta_{1,c}}} \right) \right] \\ &\times \mathcal{W}_\phi(Z_i, z^c, z^d), \end{aligned} \quad (\text{B.12})$$

where the weight function $\mathcal{W}_\phi(Z_i, z^c, z^d)$ is a product kernel designed to accommodate the mixed nature of the data. Specifically, it takes the form $\mathcal{W}_\phi(Z_i, z^c, z^d) = K_{h_1}(Z_i^c, z^c) L(Z_i^d, z^d, \lambda_z)$, with $K_{h_1}(\cdot)$ being a continuous kernel function with bandwidth h_1 and $L(\cdot)$ being a kernel function for categorical components as defined in (3.7).

The estimation procedure follows the same iterative Newton-Raphson algorithm outlined for the purely continuous case, and thus generalizes seamlessly to the mixed-data setting. Once the localized parameters $\widehat{\theta}_0$ and $\widehat{\theta}_{1,c}$ are obtained, the estimator for the conditional probability function at a point $z = (z^c, z^d)$ is computed as

$$\widehat{p}(z) = \widehat{E}(W_i^d = 1 | Z_i = z) = \frac{1}{1 + e^{-\widehat{\theta}(z)}}, \quad (\text{B.13})$$

where $\widehat{\theta}(z) = \widehat{\theta}_0$ is the estimated local function at the target point.

B.2 Asymptotic Properties - Semiparametric Models with Discrete Endogenous Variables

In this subsection, we provide the asymptotic behavior of the estimators proposed in equations (3.12) and (3.14), which correspond to the parametric component η and the nonparametric component $m(\cdot)$, respectively.

To establish the asymptotic properties of the estimator for η in partially linear models, we need to ensure that the nonparametric part of the regression model does not distort the behavior of the parametric estimator since its estimation procedure relies on projecting out the nonparametric component. For this projection to be well-behaved, we follow Robinson (1988) and impose the following definition that enables us to ensure that the nuisance function exhibits sufficient smoothness for accurate nonparametric estimation and satisfies moment conditions that guarantee the convergence of empirical processes. Therefore, this definition ensures that the resulting parametric estimator yields reliable inference, especially root- N consistency and asymptotic normality.

Definition B.1 $\mathcal{G}_\mu^\alpha$, $\alpha > 0$, $\mu > 0$, is the class of functions $g : \mathcal{R}^d \rightarrow \mathcal{R}$ satisfying the following conditions:

- g is $(\rho - 1)$ -times partially differentiable, for $\rho - 1 \leq \mu < \rho$;
- for some $\rho_0 > 0$, $\sup_{y \in \phi_{z\rho_0}} |g(y) - g(z) - Q_g(y, z)|/|y - z|^\mu \leq G_g(z)$ for all z , where $\phi_{z\rho_0} = \{y : |y - z| < \rho_0\}$;
- $Q_g = 0$ when $\rho = 1$; Q_g is a $(\rho - 1)$ th degree homogenous polynomial in $y - z$ with coefficients the partial derivatives of g at z of orders 1 through $\rho - 1$ when $\rho > 1$;
- $g(z)$, its partial derivatives of order $\rho - 1$ and less, and $G_g(z)$ have finite α th moments.

Under Definition B.1, the functions belonging to the class $\mathcal{G}_\mu^\alpha$ are locally approximable by a Taylor expansion of order $\rho - 1$, where ρ is the smallest integer strictly greater than μ . Also, such functions must be continuously differentiable up to order $\rho - 1$ and the approximation error is required to decay at a rate governed by $\mu - \rho + 1$, thereby ensuring a form of local Hölder continuity. Additionally, the function itself, its derivatives, and the bound on the approximation error must

all have finite α th moments. To sum up, this framework generalizes classical smoothness conditions by combining local regularity with global integrability, allowing for unbounded or only weakly smooth functions.

In the following we denote $\Upsilon_{v,i} = V_i - \xi_{v,i}$ and $\Upsilon_{\Phi,i} = \Phi - \xi_{\Phi,i}$, where $\xi_{v,i} = \mathbb{E}[V_i|X_{1i}]$ and $\xi_{\Phi,i} = \mathbb{E}[\Phi|X_{1i}]$. Let $\kappa = \int k(v)^2 dv$ and $\kappa_2 = \int k(v)v^2 dv$. The following conditions are imposed to ensure the estimators meet the required smoothness, moment, and identifiability properties for valid asymptotic inference.

Assumption B.1

- (i) The sequence $\{Y_i, X_{1i}, X_{2i}, W_i^d, \epsilon_i, Z_{1i}\}_{i=1}^N$ consists of independent and identically distributed (i.i.d.) observations. The error term ϵ_i satisfies $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_\epsilon^2 < \infty$, and possesses finite fourth-order moments, i.e., $E(\epsilon_i^4) < \infty$. Also, the continuous component X_{1i} admits a probability density function $f \in \mathcal{G}_{v-1}^\infty$ and the nonparametric function is $m(\cdot) \in \mathcal{G}_v^4$, where $v \geq 2$ is an integer that denotes the degree of differentiability.
- ii) For all $p_i = p(Z_i)$, $i = 1, \dots, N$, there exist predictors $\widehat{p}_i = \widehat{p}(Z_i)$ such that $\sup_i |\widehat{p}_i - p_i| = o_p(1)$.
- (iii) $K(\cdot)$ is a product kernel and the univariate kernel $k(\cdot)$ is a bounded v th order kernel, and $k(v) = O(1/(1 + |v|)^{v+1})$.
- (vi) As $N \rightarrow \infty$, $h_2 \rightarrow 0$ and $Nh_2^{2q_{x1}} \rightarrow \infty$ and $Nh_2^{4v} \rightarrow 0$.
- (v) The following matrix function is non-singular:

$$E[(\Phi_i - E[\Phi_i|X_{1i}])(V_i - E[V_i|X_{1i}])^\top f^2(X_{1i})].$$

These assumptions are commonly adopted in semiparametric estimation frameworks and provide the theoretical foundations needed to derive consistency, asymptotic normality, and efficiency bounds for the estimator proposed for η with this

alternative approach. Assumption B.1(i) ensures that the sampling process is well-behaved and that the relevant variables exhibit sufficient regularity. Further, the smoothness requirements on $f(\cdot)$ and $m(\cdot)$ are standard in the semiparametric literature and facilitate the derivation of uniform convergence rates for the corresponding estimators. Assumption B.1(ii) ensures that the kernel estimator possesses desirable smoothing properties since the decay condition on $k(v)$ controls the tails of the kernel and reduces boundary bias. Assumption B.1(iii) imposes rate conditions on the bandwidth to balance asymptotic bias and variance, a crucial aspect for deriving the convergence rate and asymptotic normality of the estimator. Finally, Assumption B.1(iv) ensures that the model is sufficiently informative for the identification of η , which is particularly important in the presence of endogeneity.

The following theorem collects the main asymptotic features of the parametric estimator $\widehat{\eta}$ proposed in equation (3.12) for η in the second stage of the estimation procedure.

Theorem 2.1 *Let $\Upsilon_{v,i} = V_i - \xi_{v,i}$ and $\Upsilon_{\Phi,i} = \Phi - \xi_{\Phi,i}$. Under Assumption B.1, as $N \rightarrow \infty$ we get*

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \Psi_b^{-1} \Omega_b \Psi_b^{-1}),$$

provided Ψ_b is positive definite, where $\Psi_b = E(\Upsilon_{\Phi,i} \Upsilon_{v,i}^\top f_i^2)$ and $\Omega_b = \sigma_\epsilon^2 E(\Upsilon_{\Phi,i} \Upsilon_{\Phi,i}^\top f_i^4)$ and $f_i = f(X_{1i})$.

Theorem 2.1 confirms that the estimator proposed for η is $\sqrt{N}$ -consistent and asymptotically normal. This result validates the use of standard inference procedures and confidence interval construction for the parametric component of the model under standard regularity conditions. The asymptotic variance, given by $\Psi_b^{-1} \Omega_b \Psi_b^{-1}$, accounts for both the endogeneity of the discrete explanatory variable and the semiparametric structure of the model. The matrix Ψ_b captures the

strength of the instruments, while Ω_b reflects the variability in the structural disturbances.

To establish the asymptotic behavior of the nonparametric estimator $\widehat{m}(x_1)$, we introduce the following set of regularity conditions.

Assumption B.2

- (i) For some $\delta_0 > 0$, $E[|\epsilon_i|^{4+\delta_0}] < \infty$.
- (ii) As $N \rightarrow \infty$, $h_2 \rightarrow 0$ and $Nh_2^{q_{x_1}} \rightarrow \infty$.

The following theorem contains the main asymptotic properties of the nonparametric estimator $\widehat{m}(x_1)$.

Theorem 2.2 *Under Assumptions B.1 and B.2, as $N \rightarrow \infty$ we get*

$$\sqrt{Nh_2^{q_{x_1}}} \left\{ \widehat{m}(x_1) - m(x_1) - \frac{h_2^2 \kappa_2^{q_{x_1}}}{2} \text{tr}\{\mathcal{H}_m(x_1)\} \right\} \xrightarrow{d} N\left(0, \frac{\sigma_\epsilon^2 \kappa^{q_{x_1}}}{f(x_1)}\right),$$

where $\mathcal{H}_m(x_1)$ is the Hessian matrix of $m(x_1)$.

Under standard regularity conditions, the results in Theorem 2.2 corroborates that the estimator is asymptotically normal with a bias of order $O(h_2^2)$ and variance of order $O(1/Nh_2^{q_{x_1}})$, consistent with the properties of the local linear estimator as detailed in Fan & Gijbels (1996). The presence of the trace of the Hessian $\mathcal{H}_m(x_1)$ in the bias term reflects the curvature of the unknown function $m(x_1)$, while the variance depends inversely on the density $f(x_1)$ and directly on the kernel constants. Notably, the local linear estimator improves upon the bias of the local constant estimator without increasing the asymptotic variance, thereby achieving a favorable bias-variance trade-off and ensuring its efficiency in semiparametric estimation.

B.3 Empirical Illustration

B.3.1 Data

We use firm-level data from the tenth Unicredit–Capitalia survey, merged with balance-sheet records from AIDA and the Chambers of Commerce. Output, capital, and labor are measured in natural logs for 2006. Process innovation is a binary indicator equal to one if a firm reports at least one process innovation between 2004 and 2006.

The data used for the analysis are sourced from the tenth Survey on Manufacturing Firms (Indagine sulle Imprese Manifatturiere), provided by Unicredit-Capitalia, which are complemented with balance sheets sourced from either AIDA (Italian Balance Sheet Dataset of the Bureau van Dijk) or the Chambers of Commerce Registry (Unicredit2008). These data are also used in Antonioli et al. (2021). The same survey, albeit in different waves, has been largely used in the economics literature investigating firms' innovation activities (see, e.g., Parisi et al. 2006; B. H. Hall, Lotti, et al. 2009).

- **The natural logarithm of value added (y):** This measure is based on 2006 data and is reported on the balance sheet.
- **The natural logarithm of capital stock (k):** This measure uses 2006 data and is calculated by summing the value of fixed assets. It is estimated through a perpetual inventory method, considering a standard depreciation rate of 0.05 and including investments. Data for both fixed assets and investments are available from 1998 to 2006, and both are deflated using the respective aggregate price index, derived from the Italian National Statistical Office (ISTAT).
- **The natural logarithm of labor (ℓ):** This represents the number of employ-

ees reported on the balance sheet for 2006.

- **Process Innovation (*IP*):** A dummy variable that equals 1 if the firm reported implementing at least one process innovation between 2004 and 2006, and 0 otherwise. While previous studies often emphasize innovation in general (Antonioli et al. 2021), we specifically focus on process innovation. Empirical research indicates that process innovation primarily enhances productivity by reducing production costs, whereas product innovation is more likely to influence market demand (Jaumandreu & Mairesse 2010; Jaumandreu & Mairesse 2017; Jaumandreu & Lin 2018). Additionally, studies on the relationship between technological progress and employment suggest that process innovation is more strongly linked to labor displacement, while product innovation tends to have a labor-friendly effect (Vivarelli 2014; Harrison et al. 2014; Calvino & Virgillito 2018). Given these findings, we emphasize process innovation as a crucial factor in analyzing firm performance and its implications for the labor market.
- **Sectors (*S*):** Manufacturing firms are classified by 22 different sectors according to the two-digit ATECO2002 classification, which is derived from the NACE Rev.1.1 Eurostat classification.

We apply a two-step cleaning procedure to the raw dataset. First, we drop all observations with missing values in any of the variables entering the production function. Second, we trim the sample at the 2.5th and 97.5th percentiles of each variable, thereby restricting the analysis to the central 5–95% interval of the distribution and limiting the influence of outliers. After these steps, the baseline sample comprises 2924 firms. When we move to the IV specifications, the sample shrinks further because the IVs for *IP* contain additional missing values; the resulting estimation sample consists of 2257 firms.

The descriptive statistics for the main variables are reported in Table B.1.

Table B.1: Descriptive statistics

Variables	Mean	St. Dev.	Min	1st Qu.	Max
y	7.441	0.669	5.999	6.919	9.005
k	7.595	0.955	5.283	6.911	9.434
ℓ	3.479	0.643	2.398	2.944	4.836

B.4 Instrumental Variables Selection

The IVs used in the analysis are selected through a multi-step procedure based on Antonioli et al. (2021), ensuring both statistical validity and economic relevance. The initial set of potential IVs is chosen based on the KPF framework, which suggests that knowledge inputs and absorptive capacity factors serve as natural instruments. These determinants include formal and informal research activities, demand pull, technology push, and market size (Duguet 2006; Musolesi & Huiban 2010). However, given the potential endogeneity of firm-level variables, an empirical selection process is necessary. In the first step, a broad set of potential IVs is identified based on the determinants of innovation. The second step applies a backward selection procedure, regressing innovation on the candidate IVs and selecting those strongly correlated with innovation, in line with general-to-specific modeling approaches (Hendry 2000). Finally, in the third step, IV-2SLS regressions are estimated for different combinations of the selected instruments, and only the sets that satisfy strongness and validity criteria, such as the Wooldridge overidentifying restrictions test (Wooldridge 2015) and the Montiel-Pflueger test (Olea & Pflueger 2013), are retained (see Antonioli et al. (2021) for further details). Following this procedure, six sets of IVs are identified, each combining different knowledge inputs and factors of absorptive capacity. These in-

clude variables such as BetterProd (investment in product improvement), FinanIncent (receiving fiscal/financial incentives), MarkPenEU15 (market penetration in EU15), EUCompet (perceived competitiveness in the EU market), CommAgrSouthAm (commercial activity within South America), and CommAgrSouthCina (commercial activity with China). Each set represents a different combination of these variables, ensuring a balance between instrument strength and validity. Among the six sets, IV35 — combining BetterProd, MarkPenEU15, and FinanIncent — is selected as the baseline specification; full details on the robustness and validity of each IV set across alternative measures of innovation are provided in Antonioli et al. (2021).

Table B.2: Instrumental variables sets used in estimation

IV Set	Instruments
IV4	BetterProd, MarkPenEU15
IV19	MarkPenEU15, CommAgrSouthAm
IV28	BetterProd, CommAgrSouthAm, MarkPenEU15
IV31	BetterProd, Age, MarkPenEU15
IV34	BetterProd, MarkPenEU15, CommAgrSouthAm
IV35	BetterProd, MarkPenEU15, FinanIncent

Appendix C

Chapter 4

C.1 Data

The data used for the analysis are sourced from the tenth Survey on Manufacturing Firms (Indagine sulle Imprese Manifatturiere), provided by Unicredit-Capitalia, which are complemented with balance sheets sourced from either AIDA (Italian Balance Sheet Dataset of the Bureau van Dijk) or the Chambers of Commerce Registry (Unicredit2008). These data are also used in Antonioli et al. (2021). The same survey, albeit in different waves, has been largely used in the economics literature investigating firms' innovation activities (see, e.g., Parisi et al. 2006; B. H. Hall, Lotti, et al. 2009). Therefore, the present work is fully comparable to previous relevant studies.

- **The natural logarithm of value added (y):** This measure is based on 2006 data and is reported on the balance sheet.
- **The natural logarithm of capital stock (k):** This measure uses 2006 data and is calculated by summing the value of fixed assets. It is estimated through a perpetual inventory method, considering a standard depreciation rate of 0.05 and including investments. Data for both fixed assets and

investments are available from 1998 to 2006, and both are deflated using the respective aggregate price index, derived from the Italian National Statistical Office (ISTAT).

- **The natural logarithm of labor (ℓ):** This represents the number of employees reported on the balance sheet for 2006.
- **Process Innovation (IP):** A dummy variable that equals 1 if the firm reported implementing at least one process innovation between 2004 and 2006, and 0 otherwise. While previous studies often emphasize innovation in general (Antonioli et al. 2021), we specifically focus on process innovation. Empirical research indicates that process innovation primarily enhances productivity by reducing production costs, whereas product innovation is more likely to influence market demand (Jaumandreu & Mairesse 2010; Jaumandreu & Mairesse 2017; Jaumandreu & Lin 2018). Additionally, studies on the relationship between technological progress and employment suggest that process innovation is more strongly linked to labor displacement, while product innovation tends to have a labor-friendly effect (Vivarelli 2014; Harrison et al. 2014; Calvino & Virgillito 2018). Given these findings, we emphasize process innovation as a crucial factor in analyzing firm performance and its implications for the labor market.
- **Sectors (S):** Manufacturing firms are classified by 22 different sectors according to the two-digit ATECO2002 classification, which is derived from the NACE Rev.1.1 Eurostat classification.

We apply a two-step cleaning procedure to the raw dataset. First, we drop all observations with missing values in any of the relevant variables. Second, we trim the sample at the 2.5th and 97.5th percentiles of each variable, restricting

the analysis to the central 5–95% interval of the distribution and limiting the influence of outliers. After cleaning, the baseline sample comprises 2924 firms for specifications in which only capital and labour enter the production function. When process innovation is treated as endogenous, the sample reduces to 2257 firms due to additional missing values in the corresponding instruments. When labour is also treated as endogenous, the requirement of a further set of instruments introduces additional missing values, bringing the estimation sample down to 1890 firms.

The descriptive statistics of the key inputs is reported in Table C.1.

Table C.1: Descriptive statistics

Statistic	Mean	St. Dev.	Min	Max
y	7.441	0.669	5.999	9.005
k	7.595	0.955	5.283	9.434
l	3.479	0.643	2.398	4.836

C.2 Instrumental Variables Selection

The IVs used in the analysis are selected through a multi-step procedure based on Antonioli et al. (2021), ensuring both statistical validity and economic relevance. The initial set of potential IVs is chosen based on the KPF framework, which suggests that knowledge inputs and absorptive capacity factors serve as natural instruments. These determinants include formal and informal research activities, demand pull, technology push, and market size (Duguet 2006; Musolesi & Huiban 2010). However, given the potential endogeneity of firm-level variables, an empirical selection process is necessary. In the first step, a broad set of potential IVs is identified based on the determinants of innovation. The second step applies a backward selection procedure, regressing innovation on the candidate

IVs and selecting those strongly correlated with innovation, in line with general-to-specific modeling approaches (Hendry 2000). Finally, in the third step, IV-2SLS regressions are estimated for different combinations of the selected instruments, and only the sets that satisfy strongness and validity criteria, such as the Wooldridge overidentifying restrictions test (Wooldridge 2015) and the Montiel-Pflueger test (Olea & Pflueger 2013), are retained (see Antonioli et al. (2021) for further details). Following this procedure, six sets of IVs are identified, each combining different knowledge inputs and factors of absorptive capacity. These include variables such as BetterProd (investment in product improvement), FinanIncent (receiving fiscal/financial incentives), MarkPenEU15 (market penetration in EU15), EUCompet (perceived competitiveness in the EU market), CommAgrSouthAm (commercial activity within South America), and CommAgrSouthCina (commercial activity with China). Each set represents a different combination of these variables, ensuring a balance between instrument strength and validity. Among the six sets, IV35 — combining BetterProd, MarkPenEU15, and FinanIncent — is selected as the baseline specification; full details on the robustness and validity of each IV set across alternative measures of innovation are provided in Antonioli et al. (2021).

Table C.2: Instrumental variables sets used in estimation

IV Set	Instruments
IV4	BetterProd, MarkPenEU15
IV19	MarkPenEU15, CommAgrSouthAm
IV28	BetterProd, CommAgrSouthAm, MarkPenEU15
IV31	BetterProd, Age, MarkPenEU15
IV34	BetterProd, MarkPenEU15, CommAgrSouthAm
IV35	BetterProd, MarkPenEU15, FinanIncent

C.3 Heterogeneity across Sector-Level Productivity

The sector dummy coefficients $\hat{\beta}_j$, reported in detail in Table C.3 and derived from the estimates in Table 4.4, measure the conditional log-productivity differential of each sector relative to the reference sector (ATECO 14, food and beverages), holding the input bundle (k, ℓ) and IP status fixed. These intercept shifts capture permanent, sector-specific TFP heterogeneity not attributable to inputs or innovation — broadly interpretable as differences in production technology, market structure, and product quality across industries.

Two patterns emerge consistently across columns (iii) and (v). First, the ordering of $\hat{\beta}_j$ broadly aligns with the Eurostat technology-intensity classification. In column (iii), chemicals (ATECO 23, $\hat{\beta}_{23} = 0.361$), rubber and plastics (ATECO 24, $\hat{\beta}_{24} = 0.209$), and fabricated metals (ATECO 27, $\hat{\beta}_{27} = 0.161$) rank among the highest, while low-technology sectors — publishing and printing (ATECO 21, $\hat{\beta}_{21} = -0.142$), wearing apparel (ATECO 17, $\hat{\beta}_{17} = -0.089$), and wood products (ATECO 19, $\hat{\beta}_{19} = -0.078$) — record negative differentials relative to the reference, consistent with the literature on TFP dispersion across manufacturing industries (see, e.g., Doraszelski & Jaumandreu 2013). Second, in column (v), where labour endogeneity is additionally controlled for, all sector differentials turn positive and increase in magnitude.

Table C.3: Sector-level productivity differentials: PLM-NP-IV estimates

Code	Description	Tech. intensity	Endog. IP	Endog. ℓ and IP
			PLM-NP-IV (iii)	PLM-NP-IV (v)
15	Textiles	Low	−0.027	0.446
17	Wearing apparel	Low	−0.089	0.281
18	Leather and leather products	Low	0.064	0.445
19	Wood and wood products	Low	−0.078	0.257
20	Paper and paper products	Low	−0.079	0.313
21	Publishing and printing	Low	−0.142	0.228
36	Other manufacturing	Low	−0.072	0.266
22	Coke and refined petroleum	Medium-low	−0.001	0.353
25	Other non-metallic mineral prod.	Medium-low	0.018	0.433
26	Basic metals	Medium-low	0.030	0.422
30	Electrical machinery n.e.c.	Medium-low	−0.012	0.333
35	Other transport equipment	Medium-low	0.061	0.372
23	Chemicals and chemical products	Medium-high	0.361	0.747
24	Rubber and plastic products	Medium-high	0.209	0.573
27	Fabricated metal products	Medium-high	0.161	0.548
28	Machinery and equipment n.e.c.	Medium-high	0.078	0.530
31	Electrical machinery	Medium-high	0.111	0.468
34	Motor vehicles and trailers	Medium-high	−0.023	0.325
29	Office machinery and computers	High	0.196	0.589
32	Radio, TV and communication equip.	High	0.122	0.644
33	Medical and optical instruments	High	0.114	0.511
<i>Reference: ATECO 14 — Manufacture of food products and beverages (Low)</i>				
N			2,598	1,890

Notes: Estimated sector dummy coefficients $\hat{\beta}_j$ from equation (4.3), for columns (iii) and (v) of Table 4.4. Each $\hat{\beta}_j$ is the conditional log-productivity differential of sector j relative to ATECO 14, holding inputs (k, ℓ) and IP fixed. Sector dummies enter the linear part of the PLM as controls; no endogeneity is assumed for them. Column (iii): IP instrumented using first stage fitted probabilities. Column (v): labour additionally controlled via control function approach. *Tech. intensity* follows the Eurostat classification at NACE Rev. 1.1 2-digit level: *High* — above-average R&D intensity; *Medium-high* — moderately high R&D intensity; *Medium-low* — capital-intensive, low R&D content; *Low* — labour-intensive, minimal R&D activity.

Bibliography

- Acemoglu, D., U. Akcigit, H. Alp, N. Bloom, and W. Kerr (2018). Innovation, re-allocation, and growth. *American Economic Review* 108 (11), 3450–3491.
- Acemoglu, D., S. Johnson, and J. A. Robinson (Dec. 2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review* 91 (5), 1369–1401.
- Akerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007). Econometric tools for analyzing market outcomes. *Handbook Of Econometrics* 6, 4171–4276.
- Akerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica* 83 (6), 2411–2451.
- Aitchison, J., and C. G. G. Aitken (1976). Multivariate binary discrimination by the kernel method. *Biometrika* 63 (3), 413–420.
- Angrist, J., and G. Imbens (1995). *Identification and estimation of local average treatment effects*.
- Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *Journal Of Business & Economic Statistics* 19 (1), 2–28.
- Angrist, J. D., and A. B. Krueger (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal Of Economics* 106 (4), 979–1014.
- Antonioli, D., G. Gioldasis, and A. Musolesi (2021). Estimating a non-neutral production function. *Oxford Economic Papers* 73 (2), 856–878.

- Atkinson, A. B., and J. E. Stiglitz (1969). A new view of technological change. *The Economic Journal* 79 (315), 573–578.
- Blundell, R., and M. C. Dias (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal Of Human Resources* 44 (3), 565–640.
- Blundell, R., and J. L. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. *Econometric Society Monographs* 36, 312–357.
- Blundell, R. W., and J. L. Powell (2004). Endogeneity in semiparametric binary response models. *Review Of Economic Studies* 71 (3), 655–679.
- Bogliacino, F., M. Piva, and M. Vivarelli (2012). R&D and employment: An application of the LSDVC estimator using European microdata. *Economics Letters* 116 (1), 56–59.
- Calvino, F., and M. E. Virgillito (2018). The innovation-employment nexus: a critical survey of theory and empirics. *Journal Of Economic Surveys* 32 (1), 83–117.
- Card, D. (1999). “Chapter 30 - The Causal Effect of Education on Earnings”. *Handbook of Labor Economics*. Ed. by O. C. Ashenfelter, and D. Card. Vol. 3. Elsevier, 1801–1863.
- Cardot, H., and A. Musolesi (2020). Modeling temporal treatment effects with zero inflated semi-parametric regression models: The case of local development policies in France. *Econometric Reviews* 39 (2), 135–157.
- Centorrino, S., F. Feve, and J.-P. Florens (2017). Additive Nonparametric Instrumental Regressions: A Guide to Implementation. *Journal Of Econometric Methods* 6 (1), 20150010.
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook Of Econometrics* 6, 5549–5632.
- Christensen, L. R., D. W. Jorgenson, and L. J. Lau (1973). Transcendental Logarithmic Production Frontiers. *The Review Of Economics And Statistics* 55 (1), 28–45.

- Chudnovsky, D., A. López, and G. Pupato (2006). Innovation and Productivity in Developing Countries: A Study of Argentine Manufacturing Firms' Behavior (1992–2001). *Research Policy* 35 (2), 266–288.
- Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal Of The American Statistical Association* 74 (368), 829–836.
- Cobb, C. W., and P. H. Douglas (1928). A Theory of Production. *American Economic Review* 18 (Supplement), 139–165.
- Crépon, B., E. Duguet, and J. Mairessec (1998). Research, innovation and productivity [ty: an econometric analysis at the firm level. *Economics Of Innovation And New Technology* 7 (2), 115–158.
- Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011). Nonparametric instrumental regression. *Econometrica* 79 (5), 1541–1565.
- Das, M. (2005). Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal Of Econometrics* 124 (2), 335–361.
- Delgado, M. S., and C. F. Parmeter (2014). A simple estimator for partial linear regression with endogenous nonparametric variables. *Economics Letters* 124 (1), 100–103.
- Doraszelski, U., and J. Jaumandreu (2013). R & D and productivity: Estimating endogenous productivity. *Review Of Economic Studies* 80 (4), 1338–1383.
- Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal Of Economics/Revue Canadienne d'économique* 36 (4), 767–808.
- Duguet, E. (2006). Innovation height, spillovers and TFP growth at the firm level: Evidence from French manufacturing. *Economics Of Innovation And New Technology* 15 (4-5), 415–442.
- Edwards, L. N. (1978). An empirical analysis of compulsory schooling legislation, 1940-1960. *The Journal Of Law And Economics* 21 (1), 203–222.

- Engle, R. F., C. W. Granger, J. Rice, and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal Of The American Statistical Association* 81 (394), 310–320.
- Fan, J. (1992). Design-adaptive Nonparametric Regression. *Journal Of The American Statistical Association* 87 (420), 998–1004.
- Fan, J., and I. Gijbels (1992). Variable Bandwidth and Local Linear Regression Smoothers. *The Annals Of Statistics* 20 (4), 2008–2036.
- (1996). *Local Polynomial Modeling and its Applications*. Chapman & Hall, London.
- Fan, J., N. E. Heckman, and M. P. W. and (1995). Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. *Journal Of The American Statistical Association* 90 (429), 141–150.
- Frölich, M. (2006). Non-parametric regression for binary dependent variables. *The Econometrics Journal* 9 (3), 511–540.
- Gallant, A. R. (1981). On the bias in flexible functional forms and an essentially unbiased form: the Fourier flexible form. *Journal Of Econometrics* 15 (2), 211–245.
- Gao, Q., L. Liu, and J. S. Racine (2015). A partially linear kernel estimator for categorical data. *Econometric Reviews* 34 (6-10), 959–978.
- Gourieroux, C., A. Monfort, E. Renault, and A. Trognon (1987). Generalised residuals. *Journal Of Econometrics* 34 (1), 5–32.
- Gozalo, P., and O. Linton (2000). Local nonlinear least squares: Using parametric information in nonparametric regression. *Journal Of Econometrics* 99 (1), 63–106.
- Griffith, R., E. Huergo, J. Mairesse, and B. Peters (2006). Innovation and productivity across four European countries. *Oxford Review Of Economic Policy* 22 (4), 483–498.

- Griliches, Z. (1998). "Patent statistics as economic indicators: a survey". *R&D and productivity: the econometric evidence*. University Of Chicago Press, 287–343.
- Griliches, Z., and J. Mairesse (1995). *Production functions: the search for identification*.
- Hadamard, J. (2014). *Lectures on Cauchy's problem in linear partial differential equations*. Courier Corporation.
- Hall, B. H. (2011). *Innovation and productivity*. Tech. rep. National bureau of economic research.
- Hall, B. H., F. Lotti, and J. Mairesse (2009). Innovation and productivity in SMEs: empirical evidence for Italy. *Small Business Economics* 33, 13–33.
- Hall, B. H., and J. Mairesse (1995). Exploring the relationship between R&D and productivity in French manufacturing firms. *Journal Of Econometrics* 65 (1), 263–293.
- Hall, B. H., J. Mairesse, and P. Mohnen (2010). "Measuring the returns to R&D". *Handbook of the Economics of Innovation*. Vol. 2. Elsevier, 1033–1082.
- Hall, P., and J. L. Horowitz (2005). Nonparametric Methods for Inference in the Presence of Instrumental Variables. *The Annals Of Statistics* 33 (6), 2904–2929.
- Hall, P., J. Racine, and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal Of The American Statistical Association* 99 (468), 1015–1026.
- Härdle, W., H. Liang, and J. Gao (2000). *Partially linear models*. Springer-Verlag Berlin, Germany.
- Harrison, R., J. Jaumandreu, J. Mairesse, and B. Peters (2014). Does innovation stimulate employment? A firm-level analysis using comparable micro-data from four European countries. *International Journal Of Industrial Organization* 35, 29–43.

- Hastie, T., and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall, Washington.
- Hayfield, T., and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal Of Statistical Software* 27, 1–32.
- Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* 47 (1), 153–161.
- (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* 46 (4), 931–959.
- Heckman, J. J., and R. Robb (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal Of Econometrics* 30 (1), 239–267.
- Heckman, J. J., and E. Vytlačil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica* 73 (3), 669–738.
- Henderson, D. J., and S. C. Kumbhakar (2006). Public and private capital productivity puzzle: A nonparametric approach. *Southern Economic Journal* 73 (1), 219–232.
- Henderson, D. J., and C. F. Parmeter (2015). *Applied nonparametric econometrics*. Cambridge University Press.
- Henderson, D. J., and R. R. Russell (2005). Human capital and convergence: a production-frontier approach. *International Economic Review* 46 (4), 1167–1205.
- Hendry, D. F. (2000). *Econometrics: Alchemy or Science?: Essays in Econometric Methodology*. Oxford University Press.
- Hicks, J. R. (1932). Marginal productivity and the principle of variation. *Economica* 3 (35), 79–88.
- Horowitz, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica* 79 (2), 347–394.

- Hsu, Y.-C., J.-C. Liao, and E. S. Lin (2022). Two-step series estimation and specification testing of (partially) linear models with generated regressors. *Econometric Reviews* 41 (9), 985–1007.
- Huergo, E., and J. Jaumandreu (2004). Firms’ age, process innovation and productivity growth. *International Journal Of Industrial Organization* 22 (4), 541–559.
- Imbens, G. W., and W. K. Newey (2009). Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica* 77 (5), 1481–1512.
- Ivaldi, M., N. Ladoux, H. Ossard, and M. Simioni (1996). Comparing Fourier and translog specifications of multiproduct technology: Evidence from an incomplete panel of French farmers. *Journal Of Applied Econometrics* 11 (6), 649–667.
- Jaumandreu, J., and S. Lin (2018). Prices under innovation: evidence from manufacturing firms. *CEPR Discussion Paper No. DP13146*.
- Jaumandreu, J., and J. Mairesse (2006). “Using Price and Demand Information to Identify Production Functions”. Mimeo, New York.
- (2010). *Innovation and welfare: Results from joint estimation of production and demand functions*. Tech. rep. National Bureau of Economic Research.
- (2017). Disentangling the effects of process and product innovation on cost and demand. *Economics Of Innovation And New Technology* 26 (1-2), 150–167.
- Kim, W., O. B. Linton, and N. W. Hengartner (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics* 8 (2), 278–297.

- Klein, R. W., and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal Of The Econometric Society*, 387–421.
- Kress, R. (1999). “Singular Integral Equations”. *Linear Integral Equations*. New York, NY: Springer New York, 94–124.
- Kumbhakar, S. C., B. U. Park, L. Simar, and E. G. Tsionas (2007). Nonparametric stochastic frontiers: A local maximum likelihood approach. *Journal Of Econometrics* 137 (1), 1–27.
- Kumbhakar, S. C., K. Sun, and R. Zhang (2016). Semiparametric smooth coefficient estimation of a production system. *Pacific Economic Review* 21 (4), 464–482.
- Landes, W. M. (1968). The economics of fair employment laws. *Journal Of Political Economy* 76 (4, Part 1), 507–552.
- Levinsohn, J., and A. Petrin (2003). Estimating Production Functions Using Inputs to Control for Unobservables. *The Review Of Economic Studies* 70 (2), 317–341.
- Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews* 15 (3), 261–274.
- Li, Q., and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.
- Li, Q., and T. Stengos (1996). Semiparametric estimation of partially linear panel data models. *Journal Of Econometrics* 71 (1-2), 389–397.
- Li, Q., and J. M. Wooldridge (2002). Semiparametric Estimation Of Partially Linear Models For Dependent Data With Generated Regressors. *Econometric Theory* 18 (3), 625–645.

- Linton, O., and J. P. Nielsen (1995). A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration. *Biometrika* 82 (1), 93–100.
- Loader, C. R. (1996). Local Likelihood Density Estimation. *The Annals Of Statistics* 24 (4), 1602–1618.
- Lööf, H., J. Mairesse, and P. Mohnen (2017). *CDM 20 years after*.
- Ma, S., J. S. Racine, and L. Yang (2020). Nonparametric Estimation of Marginal Effects in Regression-Spline Random Effects Models. *Econometric Reviews* 39 (8), 1–24.
- Maddala, G. S. (1992). *Introduction to Econometrics*. 2nd ed. New York: Macmillan Publishing Company.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. 3. Cambridge University Press.
- Mairesse, J., and P. Mohnen (2010). “Using innovation surveys for econometric analysis”. *Handbook of the Economics of Innovation*. Vol. 2. Elsevier, 1129–1155.
- Mairesse, J., and S. Robin (2008). Innovation and Productivity in France: A Firm-Level Analysis for Manufacturing and Services (1998–2000 and 2002–2004). Paper presented at the Symposium on Innovation at Firm Level, University of Barcelona, 1 July 2008. Mimeo, CREST-ENSAE, Paris.
- Martins-Filho, C., and F. Yao (2012). Kernel-based estimation of semiparametric regression in triangular systems. *Economics Letters* 115 (1), 24–27.
- Masso, J., and P. Vahter (2008). Technological Innovation and Productivity in Late-Transition Estonia: Econometric Evidence from Innovation Surveys. *The European Journal Of Development Research* 20 (2), 240–261.
- Mohnen, P., and B. H. Hall (2013). Innovation and productivity: An update. *Eurasian Business Review* 3 (1), 47–65.

- Musolesi, A., and J.-P. Huiban (2010). Innovation and productivity in knowledge intensive business services. *Journal Of Productivity Analysis* 34 (1), 63–81.
- Newey, W. K. (1994). Kernel Estimation of Partial Means and a General Variance Estimator. *Econometric Theory* 10 (2), 233–253.
- Newey, W. K., and J. L. Powell (2003). Instrumental Variable Estimation of Nonparametric Models. *Econometrica* 71 (5), 1565–1578.
- Newey, W. K., J. L. Powell, and F. Vella (1999). Nonparametric Estimation of Triangular Simultaneous Equations Models. *Econometrica* 67 (3), 565–603.
- Olea, J. L. M., and C. Pflueger (2013). A robust test for weak instruments. *Journal Of Business & Economic Statistics* 31 (3), 358–369.
- Olley, G. S., and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica* 64 (6), 1263–1297.
- Pagan, A., and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press, Cambridge.
- Pakes, A., and Z. Griliches (1984). “Patents and R&D at the Firm Level: A First Look”. *R&D, Patents, and Productivity*. National Bureau Of Economic Research, Inc, 55–72.
- Parisi, M. L., F. Schiantarelli, and A. Sembenelli (2006). Productivity, innovation and R&D: Micro evidence for Italy. *European Economic Review* 50 (8), 2037–2061.
- Pinkse, J. (2000). Nonparametric two-step regression estimation when regressors and error are dependent. *Canadian Journal Of Statistics* 28 (2), 289–300.
- Powell, J. L. (1994). Estimation of semiparametric models. *Handbook Of Econometrics* 4, 2443–2521.
- Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric Estimation of Index Coefficients. *Econometrica* 57 (6), 1403–1430.

- Qian, J., and L. Wang (2012). Estimating semiparametric panel data models by marginal integration. *Journal Of Econometrics* 167 (2), 483–493.
- Racine, J., and Q. Li (2004a). Cross-validated local linear nonparametric regression. *Statistica Sinica* 14, 485–512.
- (2004b). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal Of Econometrics* 119 (1), 99–130.
- Rivers, D., and Q. H. Vuong (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal Of Econometrics* 39 (3), 347–366.
- Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica* 56 (4), 931–954.
- Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley & Sons, Inc., New York.
- Severini, T. A., and J. G. Staniswalis (1994). Quasi-likelihood estimation in semiparametric models. *Journal Of The American Statistical Association* 89 (426), 501–511.
- Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal Of The American Statistical Association* 84 (405), 276–283.
- Stone, C. J. (1977). Consistent Nonparametric Regression. *The Annals Of Statistics* 5 (4), 595–620.
- Su, L., and A. Ullah (2008). Local polynomial estimation of nonparametric simultaneous equations models. *Journal Of Econometrics* 144 (1), 193–218.
- Syverson, C. (June 2011). What Determines Productivity? *Journal Of Economic Literature* 49 (2), 326–65.
- Tikhonov, A. N. et al. (1943). On the stability of inverse problems. *Dokl. Akad. Nauk Sssr* 39 (5), 195–198.

- Tjøstheim, D., and B. H. Auestad (1994). Nonparametric Identification of Non-linear Time Series: Projections. *Journal Of The American Statistical Association* 89 (428), 1398–1409.
- Uzawa, H. (1961). Neutral inventions and the stability of growth equilibrium. *The Review Of Economic Studies* 28 (2), 117–124.
- Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal Of Human Resources*, 127–169.
- Vivarelli, M. (2014). Innovation, employment and skills in advanced and developing countries: A survey of economic literature. *Journal Of Economic Issues* 48 (1), 123–154.
- Wooldridge, J. M. (2003). Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters* 79 (2), 185–191.
- (2010). *Econometric analysis of cross section and panel data*. MIT Press.
- (2015). Control Function Methods in Applied Econometrics. *The Journal Of Human Resources* 50 (2), 420–445.