





HILDEGART AHUMADA / M. FLORENCIA GABRIELLI  
MARCOS HERRERA / WALTER SOSA ESCUDERO

# Una nueva econometría

Automatización, big data,  
econometría espacial y estructural



SERIE **EXTENSIÓN**  
COLECCIÓN **CIENCIA Y TECNOLOGÍA**

Una nueva econometría: automatización, big data, econometría espacial y estructural /  
María F. Gabrielli... [et al.]. -1ª ed.- Bahía Blanca: Editorial de la Universidad Nacional  
del Sur. Ediuns, 2018.

205 p.; 22 x 17 cm.

ISBN 978-987-655-201-1

1. Econometría. I. Gabrielli, María F.

CDD 330.015195



Editorial de la Universidad Nacional del Sur  
Santiago del Estero 639 | B8000HZK Bahía Blanca | Argentina  
[www.ediuns.uns.edu.ar](http://www.ediuns.uns.edu.ar) | [ediuns@uns.edu.ar](mailto:ediuns@uns.edu.ar)  
Facebook: Ediuns | Twitter: EditorialUNS



**Libro  
Universitario  
Argentino**

No se permite la reproducción parcial o total, el alquiler, la transmisión o la transformación de este libro, en cualquier forma o por cualquier medio, sea electrónico o mecánico, mediante fotocopias, digitalización u otros métodos, sin el permiso previo y escrito del editor. Su infracción está penada por las Leyes n.º 11723 y 25446.

Queda hecho el depósito que establece la ley n.º 11723  
Bahía Blanca, Argentina, septiembre de 2018.

© 2018. EdiUNS.

## **Breve historia de la AAEP**

La AAEP fue fundada en 1957 por los Dres. Juan E. Alemann, Roberto Alemann, Julio Broide, Benjamín Cornejo, Aldo Ferrer, Juan J. Guaresti (h), Carlos C. Helbling, Carlos Moyano Llerena, Julio H. G. Olivera, Federico Pinedo, Oreste Popescu, Ovidio Schiopetto, Francisco Valsecchi y el Ing. Francisco García Olano.

El origen de la AAEP se remonta a sendas invitaciones cursadas por los Dres. Adolph Jöhr y Louis Baudin, a mediados de la década de los cincuenta, a los Dres. Oreste Popescu y Julio H. G. Olivera. Jöhr y Baudin, por entonces pertenecientes a la International Economic Association, sugirieron constituir una asociación representativa de los investigadores en “economía política”. La convergencia de iniciativas de los Dres. Popescu y Olivera se cristalizó, el 26 de setiembre de 1957, con la decisión de crear la AAEP. El Dr. Olivera llevó adelante la ejecución de la fundación, la que tuvo lugar formalmente el 18 de noviembre del mismo año.

La historia de la Asociación puede dividirse en tres fases. Durante la primera etapa (1957-1965) la actividad se tradujo en encuentros periódicos para la discusión de temas específicos. En 1958 se realizó la primera reunión de análisis económico. Durante este período la AAEP constituyó varios “centros regionales”. La segunda etapa (1966-1972) se caracterizó por la incorporación a la AAEP de representantes de centros e institutos de investigación. A partir de entonces, las reuniones de centros de investigación económica se realizaron en el marco de la AAEP. Se inició en 1968 la rotación de la sede y de las autoridades ejecutivas entre los principales centros. En 1972 tuvo lugar la última reunión de la AAEP organizada sobre la base de trabajos de los centros e institutos de investigación. Desde 1973 hasta el presente la AAEP se encuentra en su tercera etapa, con su sede permanente en la ciudad de Buenos Aires. La AAEP es una institución formada por investigadores y académicos en economía y que interactúa en forma directa con los mismos. El espíritu de una amplia representación institucional y regional ha quedado impreso en la actividad de la AAEP y, en especial, en la práctica de las Reuniones Anuales.

Desde su fundación, la AAEP fue presidida por Julio H. G. Olivera (1957/68), Benjamín Cornejo (1968/70), Víctor J. Elías (1970/72 y 1978/80), Miguel E. Martínez (1972/74), Horacio Núñez Miñana (1974/76), Aldo A. Arnaudo (1976/78), Rolf R. Mantel (1980/82), Mario L. Szychowski (1982/84), Ana M. Martirena Mantel (1984/86), Luisa Montuschi (1986/88), Alfredo M. Navarro (1988/90), Rinaldo Colomé (1990/92), Juan C. De Pablo (1992/94), Eusebio C. Del Rey (1994/96), Enrique A. Bour (1996/98) y José A. Delfino (1998/00), Hildegart Ahumada (2000/02), José Luis Arrufat (2002/04), Omar O. Chisari (2004/06), Alberto Porto (2006/2008), Daniel Heymann (2008/2010), Ernesto Rezk (2010/2012), Walter Sosa Escudero (2012/2014), Juan Mario Jorrat (2014/2016) y Fernando Navajas (2016/2018).

# Consejo Directivo de la AAEP

## Presidente

Fernando Navajas (FIEL)

## Presidente Electo

Jorge Streb (UCEMA)

## Secretarios Titulares

Walter Cont (FIEL)

Jorge Puig (UNLP)

## Secretarios Suplentes

Luis Blaum (UNTREF)

Silvina Elías (UNS)

## Vocales Titulares

Héctor Martínez (UNSa)

Adrián Ramos (UBA)

M. Florencia Gabrielli (UNCuyo)

Alejandro Rodríguez (UCEMA)

Lorena Garegnani (BCRA)

Ana María Cerro (UNT)

Martín Rossi (UdeSA)

Andrea Barbero (UNS)

Ricardo Descalzi (UNC)

## Vocales Suplentes

Alejandro Neme (UNSL)

Miriam Bergés (UNMdP)

Ramiro Albrieu (CEDES)

Marcelo Resico (UCA)

José Luis Pellegrini (UNR)

Leandro Arozamena (UTDT)

Florencia Barletta (UNGS)

Juan Dip (UNM)

Sonia León (UADE)





## **Prólogo**

El objetivo principal de la Asociación Argentina de Economía Política (AAEP) desde su creación a fines de los años 50 ha sido, según nuestros estatutos, la promoción del análisis económico que tiene su base en la producción científica o de investigación al más alto nivel y bajo estándares de revisión externa por pares. La AAEP ha llevado adelante esta misión a través de un nexo federal, abierto y pluralista, de departamentos de economía, institutos y centros de investigación de las universidades del país que ha sido reconocida en la obtención del Premio Konex en 2016. Basada en esta red de núcleos de investigación, la AAEP ha tenido como actividad principal su Reunión Anual, que es un verdadero ámbito de difusión de conocimiento científico en economía, en un marco de diversidad temática, metodológica, ideológica, generacional y regional. La historia de estas reuniones, plasmadas en sus Anales, es un verdadero muestrario de la diversidad y relevancia de la producción científica sobre temas de economía en la Argentina.

Desde 2004 la AAEP ha ampliado sus actividades a través de su serie de libros *Progresos en Economía*, en donde el Consejo Directivo le solicita a investigadores destacados en temas de frontera que resuman y desarrollen avances recientes sobre tópicos de economía y econometría, de modo de ofrecer a la comunidad científica, académica y profesional una visión reciente y amplia de nuevos desarrollos. La selección de estos temas obedece en parte a la necesidad o demanda que expresan nuestros socios de contar con avances o actualizaciones en temas relevantes, pero la oferta resultante de nuestros libros también surge de reconocer y difundir el trabajo de investigadores que se vienen destacando por estar en la frontera de sus respectivas áreas.

En este contexto tenemos la gran alegría de incorporar este libro de avances en econometría a la colección de *Progresos*. El mismo nos ilustra sobre diferentes avances que extienden el campo de la “métrica”, y nos invita a revisar una tendencia firme hacia la investigación cuantitativa que ahora nos abren las posibilidades de datos, métodos y poder computacional del Siglo 21. Los capítulos se presentaron en el panel sobre *Una Nueva Econometría: Automatización, Big Data, Econometría Espacial y Estructural* que tuvo lugar en San Carlos de Bariloche el 16 de Noviembre de 2017 en la LII Reunión Anual, organizada por la Sede Andina de la Universidad Nacional de Río Negro. Felicito y agradezco especialmente a los autores por la calidad de sus trabajos, pero también quisiera reconocer -en lo personal- el esfuerzo que ellos pusieron en la organización de este volumen. Sin dudas que nuestros socios, y el público interesado en general, se beneficiarán considerablemente de la lectura de este libro, que versa sobre temas que ahora emergen como determinantes de la forma de hacer investigación en economía. Estoy seguro que nuestros antecesores en la AAEP, formados con las herramientas, datos y posibilidades de cómputo del siglo pasado, estarían muy impresionados de estos avances y de lo que los mismos proyectan a futuro.

Fernando H. Navajas  
Presidente, AAEP

## Índice general

<b>Capítulo 1. Introducción</b>	<b>15</b>
<i>H. Ahumada, M. F. Gabrielli, M. Herrera y W. Sosa Escudero</i>	
<b>Capítulo 2. Selección automática de modelos econométricos</b>	<b>19</b>
<i>Hildegart Ahumada</i>	
1. Introducción . . . . .	19
2. Características del algoritmo . . . . .	20
2.1. La econometría General a Particular . . . . .	21
2.2. Un poco de historia . . . . .	23
2.3. La versión actual (2017) . . . . .	24
2.4. Algunas opciones en el proceso de selección . . . . .	28
3. Comparación con otros algoritmos de selección automática . . . . .	30
4. Saturación con Dummies y los objetivos de la econometría . . . . .	33
5. Aplicaciones prácticas . . . . .	37
6. Reflexiones finales . . . . .	41
<b>Capítulo 3. Análisis estructural de datos de subastas</b>	<b>49</b>
<i>M. Florencia Gabrielli</i>	
1. Introducción . . . . .	49
2. Modelo canónico . . . . .	54
3. Modelos de subastas, identificación y restricciones teóricas . . . . .	56
3.1. El modelo AV simétrico . . . . .	58

3.1.1. Caracterización . . . . .	58
3.1.2. Identificación . . . . .	59
3.1.3. Restricciones testeables . . . . .	61
3.2. El modelo AV asimétrico . . . . .	63
3.2.1. Caracterización de los modelos bajo análisis . . . . .	64
3.2.2. Identificación y restricciones testeables . . . . .	65
4. Estimación estructural . . . . .	66
4.1. El modelo IPV simétrico . . . . .	67
4.2. El modelo APV simétrico . . . . .	68
4.3. El modelo IPV asimétrico . . . . .	68
4.3. El modelo APV asimétrico . . . . .	69
4.5. Métodos de estimación . . . . .	70
4.5.1. Métodos directos . . . . .	71
4.5.2. Métodos indirectos . . . . .	73
5. Estimación estructural del modelo IPV simétrico . . . . .	75
5.1. Modelo teórico e identificación . . . . .	75
5.2. Modelo de estimación . . . . .	80
6. Detectando colusión dentro del Enfoque estructural . . . . .	86
6.1. El modelo de competencia (Modelo A) . . . . .	91
6.2. El modelo con colusión (Modelo B) . . . . .	95
6.3. Tests para detectar colusión . . . . .	97
7. Conclusiones . . . . .	100

**Capítulo 4. Fundamentos de econometría espacial aplicada** **107**

*Marcos Herrera*

1. Introducción . . . . .	107
2. Conceptos y elementos espaciales . . . . .	110
2.1. Sobre la naturaleza de los datos . . . . .	110
2.2. Definiendo autocorrelación espacial . . . . .	113
2.3. Importancia de la matriz de pesos espaciales . . . . .	115

3. Especificaciones espaciales alternativas . . . . .	120
3.1. Modelos espaciales de corte transversal . . . . .	121
3.2. Modelos espaciales en datos de panel . . . . .	126
3.3. Cuestiones inferenciales y asintóticas . . . . .	131
4. Interpretación de los modelos con rezago espacial endógeno . . . . .	135
5. Aplicaciones empíricas . . . . .	140
5.1. Modelo de crecimiento (Ertur y Koch, 2007) . . . . .	140
5.2. Consumo de tabaco en datos de panel . . . . .	145
6. Comentarios finales . . . . .	148

**Capítulo 5. Big data y aprendizaje automático: Ideas y desafíos para economistas** **157**

*Walter Sosa Escudero*

1. Introducción . . . . .	157
2. Aprendizaje, estadística y econometría . . . . .	159
2.1. El paradigma inferencial frecuentista en econometría . . . . .	161
2.2. Aprendizaje automático y construcción de modelos . . . . .	163
3. Regresión para la predicción . . . . .	166
3.1. Estimar vs. construir modelos . . . . .	166
3.2. Complejidad y elección de modelos . . . . .	168
3.3. Validación cruzada . . . . .	169
3.4. Regularización: LASSO y ridge . . . . .	172
4. Clasificación . . . . .	178
4.1. Riesgo, probabilidades y clasificación . . . . .	178
4.2. Regresión logística y alternativas . . . . .	179
4.3. Categorías múltiples . . . . .	181
4.4. Árboles . . . . .	182
5. Aprendizaje no supervisado: clusters y dimensionalidad . . . . .	186
5.1. Clusters . . . . .	187
5.2. Reducción de dimensionalidad . . . . .	190

6. Epílogo: la lenta introducción de big data en la economía . . . . .	193
7. Bibliografía sugerida . . . . .	197
<b>Sobre los autores</b>	<b>203</b>

## **Capítulo 1. Introducción**

H. Ahumada, M. F. Gabrielli, M. Herrera y W. Sosa Escudero

La serie de Progresos en Economía de la AAEP tiene la intención de informar a los asociados y al público en general de los avances recientes en un campo de análisis económico específico, de modo de poner al alcance las nuevas tendencias y aportes de frontera. Este volumen es el segundo que contiene a la econometría como tema central. El auge de nuevas técnicas de reciente desarrollo y su pronunciado avance han generado la necesidad de este volumen con la idea de presentar algunas de las áreas más prolíferas.

Los cuatro trabajos que se compilan en este libro fueron presentados en el Panel de Progresos en Econometría coordinado por Walter Sosa Escudero durante la LII Reunión Anual de la Asociación Argentina de Economía Política, el 16 de noviembre de 2017 en la Universidad Nacional de Río Negro. El objetivo de este panel fue reunir a un grupo de destacados econométricos argentinos para discutir los desarrollos más recientes que la Econometría ha ido incorporando.

Hildegart Ahumada, en el segundo capítulo, se focaliza en el uso de nuevos algoritmos, y la adaptación de algunos ya existentes, para la selección

automática de modelos econométricos. Estos algoritmos han sido en general enfocados a la predicción o siendo inespecíficos en cuanto al objetivo y solo sintetizando una colección de modelos alternativos. Según su opinión, un algoritmo a destacar es *Autometrics*, ya que automatiza lo que haría un econometrista cuando realiza una modelación siguiendo una determinada metodología, en este caso la denominada general a particular, para alcanzar los distintos objetivos de modelación econométrica: inferencia, pronóstico y evaluación de políticas. Así, dicho algoritmo ayuda al investigador en la búsqueda de estimaciones robustas, evaluación de constancia de parámetros y análisis de exogeneidad, respectivamente. El capítulo describe el algoritmo, resumiendo y discutiendo los avances en la modelación obtenidos con él, enfatizando los aspectos distintivos referidos a la Saturación con dummies. También se compara *Autometrics* con otros algoritmos alternativos. Dentro de esta discusión se incluyen algunas aplicaciones interesantes, destacando potencialidades y limitaciones.

M. Florencia Gabrielli presenta una revisión de la literatura relevante para el análisis de datos de subastas desde la perspectiva del enfoque estructural que destaca la estrecha vinculación entre el modelo teórico subyacente y el modelo econométrico resultante. El análisis de subastas es un área de investigación activa en economía tanto desde un punto de vista teórico como empírico. Los mecanismos de subastas para asignar bienes y servicios son ampliamente utilizados. Esto ha generado importantes bases de datos que se utilizan para analizar estos mercados. En particular en este capítulo, la autora comienza analizando el modelo canónico que establece un marco de referencia unificado para luego desarrollar desde el enfoque estructural el marco que permite caracterizar, identificar y establecer restricciones testeables de los distintos modelos teóricos. El modelo más general de subastas se desarrolla a partir del Paradigma de Valuaciones Afiliadas que distingue dos casos polares, las Valuaciones Privadas y las Valuaciones Comunes. A su vez, cada modelo puede



clasificarse en simétrico o asimétrico según la naturaleza de sus participantes. Cada uno de estos casos conduce a distintos modelos de subastas. El capítulo luego presenta diferentes estrategias y métodos de estimación. Los métodos no paramétricos o semi-paramétricos se discuten de manera especial ya que son ampliamente utilizados en este contexto. Finalmente, el capítulo explora avances recientes sobre cómo incorporar comportamientos no cooperativos (por ejemplo, colusión) en los modelos y cómo pueden adaptarse las estrategias econométricas.

Marcos Herrera hace una reseña sobre la econometría espacial, rama econométrica dedicada al análisis de *efectos espaciales* en modelos de regresión en datos de corte transversal y de panel. Tradicionalmente, los modelos que incorporaban explícitamente el espacio geográfico estaban confinados a temas de ciencia regional, economía geográfica o urbana. Sin embargo, el actual entendimiento del espacio no está restringido a cuestiones geográficas ampliando el uso de las técnicas al espacio económico y social. La diversidad de aplicaciones que utilizan herramientas de econometría espacial incluye temas de redes sociales, criminología, finanzas públicas, ciencia política, psicología, salud, educación, historia, transporte, marketing, entre muchos otros. Herrera revisa y organiza los desarrollos más consolidados de econometría espacial, enfatizando un rango de tópicos relativos a la especificación e interpretación de los efectos espaciales centrandó la discusión en modelos con variable dependiente continua que buscan incorporar dependencia espacial. El capítulo desarrolla los principales modelos espaciales en corte transversal y en datos de panel, discutiendo la interpretación de los modelos que incluyen efectos espaciales endógenos. La discusión se ilustra mediante dos ejemplos empíricos de corte transversal y panel de datos.

Walter Sosa Escudero resume las ideas asociadas a *big data*, aprendizaje automático, minería de datos, etc. que traen una propuesta concreta de revisión sustancial de los paradigmas metodológicos de la ciencia, a la luz del virtual

“diluvio de datos” provocado por las redes sociales y el uso de dispositivos interconectados. Sosa Escudero revisa algunas ideas básicas del paradigma de *big data* y aprendizaje automático con el doble propósito de diseminar el uso de estas tecnologías, discutir sus ventajas y desventajas y reflexionar sobre las razones que explican la cautela imperante en economía. El contenido del capítulo cubre una *mínima* cantidad de material, suficiente para apreciar las principales diferencias entre el enfoque de aprendizaje automático y el clásico de la econometría, observar las principales desventajas y ventajas y, fundamentalmente, motivar e iniciar al interesado en esta literatura, priorizando intuiciones y conceptos por sobre tecnicismos. Se exploran algunas ideas básicas (construcción de modelos, complejidad, validación cruzada, etc.) dentro de un contexto familiar para los econométricos: el modelo de regresión lineal. La idea se extiende en la dirección del problema de clasificación y discute algunas estrategias de *aprendizaje no supervisado*, como la construcción de clusters o el problema de reducción de la dimensionalidad mediante diferentes ejemplos. Finalmente se ofrece una breve guía de lectura para los interesados en ahondar en esta temática.

Estas cuatro áreas econométricas con certeza no son las únicas recientemente desarrolladas. El progreso de la econometría surge por diferentes avances dentro de la ciencia económica como también por la interacción con otras disciplinas. Esto genera la dificultad de resumir en un volumen único todo este progreso. Las subjetividades de los autores, las limitaciones del tiempo y del espacio también conspiran contra desarrollos más enciclopédicos de los temas aquí presentados. Sin embargo, los autores esperan que la calidad y el interés de los trabajos incluidos en este libro sirvan como un paliativo de las mencionadas limitaciones. Los autores agradecen a Fernando Navajas y al Consejo Directivo de la AAEP por la invitación a preparar este trabajo.

## Capítulo 2. Selección automática de modelos econométricos

Hildegart Ahumada\*

### 1. Introducción

En los últimos años hemos sido testigos del desarrollo de innovaciones tecnológicas hasta hace poco impensable en cuanto a su velocidad y a la extensión a distintos campos. Asimismo, estamos presenciando el uso de nuevos algoritmos, o la adaptación de algunos ya existentes, aplicados a diversos conjuntos de información, frecuentemente a *big data*. Estos algoritmos han sido en general enfocados a un solo objetivo, como el de predicción (teniendo en su mayoría como función de pérdida, parcial o totalmente, la minimización de los errores cuadráticos) o siendo inespecíficos en cuanto al objetivo y solo sintetizando una colección de modelos alternativos.

---

\* Agradezco los comentarios y sugerencias de Magdalena Cornejo y Lucía Freira a una versión anterior, errores y omisiones son de mi responsabilidad.

Sin embargo, uno podría preguntarse, alternativamente, si habría algún algoritmo que “imite” la función de un econometrista siguiendo un determinado enfoque metodológico. *Autometrics* es un algoritmo<sup>1</sup>, según mi conocimiento el único, que automatiza lo que haría un econometrista cuando realiza una modelación siguiendo una determinada metodología, en este caso la denominada *general a particular* que, en su aplicación, se orientará a los distintos objetivos de la modelación econométrica: inferencia, pronóstico y evaluación de políticas. Para obtener dichos objetivos este algoritmo ayudará al investigador en la búsqueda de estimaciones robustas, evaluación de constancia de parámetros y análisis de exogeneidad, respectivamente.

Este trabajo se focaliza en dicho algoritmo, resumiendo y discutiendo los avances en la modelación obtenidos con él. Las secciones 2 y 3, respectivamente, describen *Autometrics* en el marco de la metodología *general a particular* y lo comparan con otros algoritmos usados. Posteriormente en la sección 4 se pondrá un mayor énfasis en los aspectos distintivos de *Autometrics* referidos a la *Saturación con dummies* y los diferentes objetivos de la modelación econométrica. La sección 5 contribuye a difundir algunas aplicaciones interesantes que se han realizado gracias a este algoritmo. La sección 6 concluye con algunas reflexiones sobre potencialidades y limitaciones.

## 2. Características del algoritmo

Para comenzar trataremos de explicar qué es *Autometrics*. Brevemente, podemos definirlo como un algoritmo que automatiza la metodología de

---

<sup>1</sup>Se encuentra disponible en el *software* Oxmetrics (Doornik y Hendry, 2013) y en R (Sucarrat et al., 2017).

modelación econométrica conocida como *general a particular* y cuyos pilares resumiremos a continuación.

## 2.1. La econometría General a Particular

Si bien se originó en los desarrollos de la econometría de series de tiempo en la London School of Economics (como se refiere a ella, por ejemplo, Favero, 2001, capítulo 3) y se la asocia con la metodología propuesta por David Hendry y sus coautores (como la identifica, Gilbert, 1986) para este tipo de datos, actualmente *General to Specific (Gets)*<sup>2</sup> es un enfoque que engloba un conjunto de metodologías que tienen en común criterios básicos de la modelación econométrica. Un resumen de estos criterios sería el siguiente:

- La utilización tanto de teoría económica como de los mismos datos a analizar como fuentes de información para postular los potenciales modelos.
- En el casi inevitable *trade-off* entre sesgo (Variables Omitidas) vs Varianza (Variables Redundantes) se tratará de evitar el primero de los problemas cuando estimamos inicialmente los coeficientes de un modelo econométrico. Así, el objetivo es considerar en la medida de lo posible, todas las colinealidades de forma de evitar inicialmente el sesgo e inconsistencia de los estimadores. Eso llevará a tener como punto de partida un modelo general irrestricto (GUM, por sus siglas en inglés) lo más “general” posible.

---

<sup>2</sup>Una colección de trabajos sobre *Gets* puede consultarse en Campos et al. (2005) y un resumen de su evolución en Ahumada (2007).

- La reducción del GUM a más modelos simples (parsimoniosos)<sup>3</sup> e interpretables en términos de teoría económica, medición de los datos, etc.
- La congruencia del GUM con el proceso generador de los datos local (LDGP, por sus siglas en inglés). Hablamos de un DGP *local* porque consideramos el subconjunto de variables del DGP, que serían necesarias para capturar el comportamiento económico relevante y no necesariamente del verdadero DGP, el cual podría implicar variables desconocidas. La evaluación de congruencia, para verificar si contamos con una aproximación satisfactoria al LDGP, nos llevará a la aplicación de tests de falta de especificación, es decir tests de diagnósticos para que la evidencia no rechace los supuestos necesarios para inferencia, como la ausencia de heteroscedasticidad y autocorrelación, entre otros. Así intentaremos asegurarnos que partimos de un modelo que es la mejor opción entre las disponibles.
- La validez del modelo condicional. Una de formas más usadas para simplificar la modelización es suponer modelos uniecuacionales donde la variable que intentamos explicar sea función de un conjunto de variables explicativas. Necesitaremos la exogeneidad de las variables explicativas, para cuya evaluación usaremos la estimación por variables instrumentales (IVE) o la estimación por sistemas.
- En el contexto de series temporales (y panel) la inclusión de variables integradas (ej.  $I(1)$ ) nos llevará a su reparametrización como variables estacionarias ( $I(0)$ ) a través de su diferenciación y su combinación lineal (como en el caso de modelo de *Error o Equilibrium Correction*).
- La robustez de los modelos con respecto a la posible presencia de valores extremos y quiebres.

---

<sup>3</sup>Un pionero en esta línea es Anderson (1962) quien propone la simplificación secuencial del orden de un polinomio en una regresión mostrando que, en general, la búsqueda de lo particular a general no es óptima, extendiéndolo luego a la elección de los rezagos.

- La consideración del concepto de No Estacionariedad en sentido amplio: la modelización de cambios estructurales (tanto en el tiempo como debido a diferentes unidades de observación).
- La evaluación de la posibilidad de No Linealidades.

La reducción del GUM y, en particular, los últimos tres criterios, como veremos más adelante, han podido ser reconsiderados y simplificados gracias a los avances en la selección automática a través de *Autometrics*.

## 2.2. Un poco de historia

Antes de describir el algoritmo, vale la pena señalar que la versión actual es en realidad la tercera generación de una familia de algoritmos.

Asimismo, su desarrollo es un excelente ejemplo de cómo la superación de críticas puede llevar al progreso científico. En este sentido, el algoritmo fue evolucionando desde sus primeros orígenes asociados al trabajo de Hoover y Perez (1999), quienes propusieron una búsqueda de múltiples sendas para implementar *Gets*. El procedimiento sugerido por estos autores surgió, en realidad, como una respuesta al estudio de Lovell (1983) quien investigó *data mining* a través de simulaciones que cuestionaban la capacidad de diferentes enfoques (*Stepwise forward and backward regressions* en función del  $R^2$ , estadísticos  $t$ , etc.) para recobrar el *DGP* que había supuesto (0-5 verdaderos regresores escondidos en 40 variables). En cambio, las simulaciones de Monte Carlo de Hoover y Perez, mostraron una sustancial mejora de los resultados obtenidos por *Gets*. Ellos, a diferencia de Lovell, usaron una búsqueda de

múltiple sendas y realizaron tests de diagnóstico, además de utilizar tamaños de muestras más apropiados.

Los resultados de Hoover y Perez llevaron a Hendry y Krolzig a desarrollar *PcGets*<sup>4</sup> el antecesor más directo de *Autometrics*. Sus principales rasgos distintivos fueron, por una parte, una sistematización de la búsqueda de acuerdo con los ordenamientos dados por los estadísticos  $t$  al cuadrado. Por otra parte, no pusieron restricciones al número de sendas, cada senda se iniciaba con una variable no significativa, mientras que Hoover y Perez solo analizaron diez. En los dos casos se hizo hincapié en el uso de tests de diagnósticos.

### 2.3. La versión actual (2017)

Continuando con la evolución, *Autometrics* (Doornik, 2009; Hendry y Doornik 2014) se diferenció principalmente de *PcGets*, porque utiliza para la selección automática de variables un procedimiento de *búsqueda de árbol eficiente*, en lugar de sendas múltiples, de acuerdo con los estadísticos  $t$  al cuadrado ordenados de mayor a menor. Por lo tanto, para el caso que estamos más acostumbrados de un mayor número de observaciones que variables ( $T > K$ ), *Autometrics* selecciona las variables relevantes a partir del modelo irrestricto GUM, provisto por el econometrista y que se supone como una buena representación del LDGP de acuerdo con los tests de diagnósticos. De las distintas combinaciones ordenadas en el árbol según los mayores  $t$  cuadrados y

---

<sup>4</sup>Una interesante aplicación de este algoritmo puede verse en Hendry y Krolzig (2004) donde muestran como los dos millones de ecuaciones del conocido estudio de los modelos de crecimiento de Sala-i-Martin pueden reducirse a sola una regression (*we run one regression* partiendo del apropiado modelo irrestricto que englobe los numerosos casos y seleccionando por *PcGets*).



a partir de un nivel de significancia, dado también por el econometrista, quedan seleccionadas las variables explicativas relevantes.

Otra diferencia con la versión anterior, que implica mayor rapidez en la aplicación, es que no se evalúan los tests de diagnósticos en cada etapa de reducción de variables. Como suponemos que el GUM es congruente, si los tests fallan en el modelo terminal, el procedimiento vuelve (*back-on-track*) hacia modelos más generales, en dirección al GUM original. También, este algoritmo alcanza mayor rapidez a través de eliminar permanentemente variables muy insignificantes (*Chopping*) y por grupos (*Bunching*). Asimismo, cuando una variable no puede ser removida se anula ese nodo y las siguientes ramas se ignoran (*Pruning*).

Cuando el algoritmo completa la búsqueda pueden existir más de un modelo terminal y, en ese caso, selecciona uno de ellos siguiendo el criterio de Schwarz, aunque existe la posibilidad de elegir otros criterios de acuerdo con la penalización elegida por el investigador, según sus objetivos en la modelización<sup>5</sup>.

Un aspecto importante a señalar es que una vez que se consideran todas las ramas del árbol, las variables que quedan dentro y fuera del modelo se definen como una *decisión de un solo corte (1-Cut-Off)*, dado un determinado nivel de significación. Por ejemplo, si el nivel de significancia es de 5% se incluirán las variables pertenecientes a las ramas del diagrama de árbol con estadísticos  $t^2$  iguales o superiores a  $1,96^2$  para una muestra grande. Así, por lo tanto, se evita la cuestión conocida como el problema de *Repeated Testing*. Dicha cuestión referida a la selección de variables de acuerdo con los estadísticos  $t$ , consistiría en el problema de calcular incorrectamente el error de *tipo I*, ya que si computamos, asumiendo independencia, la probabilidad de que la primera variable seleccionada sea no significativa por la probabilidad de que la

---

<sup>5</sup>El investigador puede incluso seleccionar varios de estos modelos para combinarlos si el objetivo fuese, por ejemplo, el de pronóstico.

segunda lo sea, dado que la primera fue no significativa y así sucesivamente, el error de *tipo I* resultante sería  $(1 - \alpha)^K$ , digamos igual a 0,87 para  $K = 40$  variables partiendo del valor tradicional de  $\alpha = 0,05$ . Esta supuesta corrección del error de *tipo I* asociada con el proceso de *Repeated Testing* no aparece si la selección de variables se corresponde con decisiones de un solo corte, como la anteriormente comentada.

Asimismo, debe señalarse que en realidad lo que importa para la selección de variables es el número promedio de variables irrelevantes retenidas en el modelo. En el ejemplo anterior, sería igual a  $\alpha.K = 2$ , valor que podría reducirse a 0,41 variables irrelevantes retenidas en promedio si usamos  $\alpha = 0,01$ , un nivel de significancia algo menor<sup>6</sup>.

En consecuencia, se podrían retener pocas variables irrelevantes con valores apropiados de  $\alpha$  (que también dependen de  $T$ , el número de observaciones). Si quisieramos que (en promedio) solo una de las  $K$  variables del modelo fuese irrelevante, entonces usaríamos  $\alpha = \frac{1}{K}$ . También si  $T$ , el número de observaciones fuese muy grande, tendríamos demasiada evidencia en contra de la hipótesis nula (Hendry, 1995) y deberíamos usar un nivel de significancia menor. Un nivel de significancia que se recomienda cuando se incluyen  $K$  variables es  $\alpha = \min(K, T, 1\%)$ . Este nivel parece más exigente que el tradicional 5%, y lleva a preguntarse por la potencia del estadístico  $t$  en estos casos. Sin embargo, se puede notar que el rechazo de la hipótesis nula depende también de los valores del estadístico  $t$  en la hipótesis alternativa y si las influencias son sustantivas, habrá poca pérdida de potencia<sup>7</sup>.

---

<sup>6</sup>Pensemos en la distribución Binomial para obtener el valor esperado de la retención de  $K$  variables irrelevantes asumiendo que son ortogonales.

<sup>7</sup>La potencia depende, en realidad, de la no centralidad del estadístico  $t$  en la muestra. Suponiendo la aproximación a la Normal con varianza unitaria, si el parámetro de no centralidad es 2 y usamos el tradicional 5%, la potencia es de 0,5 pero si es 4 es de 0,92 aunque usemos el más exigente 1%, ver Hendry y Doornik (2014, capítulo 10).

En síntesis el rol del econometrista sería el de formular el modelo inicial (*GUM*) según la información proveniente de teoría económica, de modelos rivales, etc. de forma que si determina el vector de variables explicativas  $\mathbf{x}_t$  para modelar  $y_t$ ,<sup>8</sup> el modelo puede generalizarse incluyendo  $l$  rezagos y términos no lineales de las variables básicas del tipo  $g(\mathbf{x}_{t-l})$  además de componentes determinísticos, variables indicadoras para valores extremos, quiebres y tendencia denotados por el vector  $\mathbf{d}_t$ ,

$$y_t = \sum_{j=0}^l \beta' \mathbf{x}_{t-j} + \sum_{j=0}^l \gamma' g(\mathbf{x}_{t-j}) + \delta' \mathbf{d}_t + \sum_{j=1}^l \phi' y_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim IN(0, \sigma_\varepsilon^2). \quad (1)$$

A esta ecuación se verificará su congruencia, es decir la validez empírica del supuesto de  $\varepsilon_t \sim IN(0, \sigma_\varepsilon^2)$ . A partir de allí el algoritmo seleccionará el(los) modelo(s) más parsimonioso(s) de acuerdo con un nivel de significancia que el econometrista debe también decidir. Esta última decisión es aún una de las cuestiones más críticas como veremos posteriormente.

Con respecto a la cuestión de utilizar variables integradas, debe notarse que esto no dificultaría la aplicación de *Autometrics* ya que el modelo en niveles (o logaritmos de ellos) puede reescribirse en diferencias y combinaciones lineales que sean estacionarias si las variables del modelo están cointegradas (como lo han analizado Sims et al., 1990)<sup>9</sup>, solo teniendo en cuenta que el nivel de la variable explicada rezagada (que daría origen al coeficiente de ajuste en un modelo de *Corrección de Errores*) podría tener una distribución no estándar (si no hubiera cointegración). Es decir, como ejemplo, si consideramos un caso especial de la Ecuación (1), con una sola variable explicativa integrada de orden uno al igual que  $y_t$  y un solo rezago (lineal y sin componentes determinísticos),

$$y_t = \beta_1 x_t + \beta_2 x_{t-1} + \phi y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim IN(0, \sigma_\varepsilon^2). \quad (2)$$

---

<sup>8</sup>El subíndice  $t$  no implica que deberá restringirse solo a modelos de series temporales como veremos en las aplicaciones de la sección 5.

<sup>9</sup>Una exposición simplificada de esta cuestión puede verse en Ahumada (2006).

Puede reescribirse en niveles y diferencias como,

$$\Delta y_t = \beta_1 \Delta x_t - (1 - \phi)y_{t-1} + (\beta_1 + \beta_2)x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim IN(0, \sigma_\varepsilon^2). \quad (3)$$

En el caso de variables integradas de orden uno, solo la distribución de  $\phi$  podría ser no estándar si las variables no cointegran. Sin embargo, si hay cointegración entre las variables del modelo, la combinación lineal de ellas será estacionaria implicando un modelo de *Corrección de Errores*, donde  $1 - \phi$  es el coeficiente de ajuste a la relación de cointegración, siendo el coeficiente de largo plazo  $\frac{(\beta_1 + \beta_2)}{1 - \phi}$ . Dadas la características de muchas series económicas, el valor  $\phi$  será lo suficientemente alto en (2) para ser retenido durante la selección y posteriormente, se podrá evaluar la posibilidad de no cointegración<sup>10</sup>.

Finalmente, podemos destacar algunas opciones que puede utilizar un investigador en el proceso de selección.

## 2.4. Algunas opciones en el proceso de selección

Entre ellas destacaremos las siguientes:

- Mantener una o más **variables fijas** durante la selección, según lo considere el investigador. Por ejemplo, mantener alguna variable específica del modelo económico o la constante.
- Estimar los modelos por **Variables Instrumentales** en lugar de mínimos cuadrados ordinarios (MCO) cuando algunas de las explicativas no pueden suponerse exógenas.
- Utilizar durante la selección **Errores Estándares (SE) Robustos** a heteroscedasticidad y/o autocorrelación. Dado que se busca la

---

<sup>10</sup>Para evaluar la significatividad de  $1 - \phi$  pueden usarse los valores críticos propuestos por Ericsson y Mackinnon (2002), entre otros.

congruencia del modelo inicial, el uso de estos errores no serían necesarios y en consecuencia, no sería una opción recomendada, excepto en algunos casos. Por ejemplo, si por su naturaleza los errores están necesariamente autocorrelacionados, como en el caso de los errores de pronóstico a más de un período adelante, se deberían utilizar SE robustos para la evaluación de los determinantes de su sesgo. También los SE robustos a heteroscedasticidad y/o autocorrelación se usarían cuando se trabaje con subconjuntos de las potenciales variables de un modelo, como explicaremos en el siguiente caso.

- Selección de variables para **el caso de mayor número de variables explicativas que observaciones** ( $K > T$ ), aunque siempre asumiendo que  $K \ll T$  en el *DGP*. El caso de ( $K > T$ ) ocurrirá frecuentemente cuando queremos modelar conjuntamente con las variables explicativas, sus rezagos, los valores extremos, quiebres y potenciales no linealidades del modelo. En este caso, *Autometrics* realiza la partición en grupos, lo más grandes posibles para evitar, en la medida de lo posible, perder las colinealidades pero tal que en cada uno  $K \ll T$ . Se tratará que, con estos bloques de variables, los modelos sean congruentes aunque, al trabajar con subconjuntos de información pueden no serlos y en consecuencia, se deberán usar SE robustos. Así se estimarán todas las combinaciones de los grupos para seleccionar potenciales variables, inicialmente con un nivel de significancia grande para retener de los diferentes bloques un mayor número de variables, pero usando un nivel de significancia más pequeño en la elección final. La unión de las variables seleccionadas formará el GUM, como en el caso tradicional<sup>11</sup>.

Dentro de este último caso puede considerarse el enfoque denominado *Saturación con Dummies*, el cual debido debido a su originalidad y relevancia

<sup>11</sup>La posibilidad de poder seleccionar modelos con más variables que observaciones resultan de gran utilidad para *big data* como sugieren Doornik y Hendry (2015).

para la práctica econométrica será analizado en la Sección 4, luego de realizar brevemente una comparación con otros procedimientos de selección.

### 3. Comparación con otros algoritmos de selección automática

Esta sección resume brevemente algunas alternativas disponibles para la selección de variables tratando de enfatizar en qué aspectos se distingue *Autometrics*. Un análisis más detallado puede encontrarse en Hendry y Doornik (2014, capítulo 17) donde también se reportan los resultados comparativos de diferentes simulaciones.

Entre los métodos más difundidos de selección automática encontramos:

- **STEP WISE** Este método para estimar regresiones automáticamente se basa en que en cada *step* una variable es agregada (*forward selection*) o retirada (*backward selection*) del modelo de acuerdo con la correlación con la variable a explicar, utilizando criterios como los estadísticos  $t$ ,  $F$ , el  $R^2$  o  $R^2$  ajustado, etc. Si bien, por lo tanto, este tipo de búsqueda puede ser tanto particular a general como general a particular, se usan solo criterios de bondad del ajuste y en ningún caso se evalúan los tests de diagnóstico. Además, cuando la búsqueda es de lo particular a lo general, al no controlar por todas las posibles colinealidades, puede dejar afuera variables relevantes del modelo.
- **RETINA** (*Relevant Transformation of the Inputs Network Approach*, código en Gauss y Matlab, Perez-Amaral et al., 2003). Usando transformaciones no lineales (cuadrados, inversas, multiplicaciones de variables) de los predictores selecciona entre los posibles modelos de acuerdo con la mejor capacidad de pronóstico (*out-of-sample*),

controlando por colinealidades. El orden de las variables seleccionadas depende de la correlación muestral con la variable dependiente y que satisfaga la condición de que esté debajo de un nivel de colinealidad dado con las variables ya incluidas. Su objetivo puede resumirse como el de buscar un conjunto parsimonioso de variables (transformadas) siguiendo un enfoque particular a general.

- **LASSO** (*Least Absolute Shrinkage and Selection Approach*, Tibshirani, 1996 y Efron et al., 2004) actualmente muy en uso para *big data* y cuando se busca limitar el número de (muchos) controles poco significativos<sup>12</sup>. Su objetivo es minimizar la suma de errores al cuadrado, al igual que MCO, pero sujeto a que el valor absoluto de los coeficientes estimados  $b_j$  sean menores que un valor especificado  $s$ . Debe notarse que si  $s$  es lo suficientemente grande los  $b_j$  coincidirán con los obtenidos por MCO. Sin embargo, valores más pequeños de  $s$  implican *shrinkage* de los  $b_j$  obtenidos por MCO, pudiendo algunos ser cero. La elección de  $s$  ha tendido a realizarse por *cross-validation (out-of-sample)*. Actualmente se realiza a través del algoritmo LARS (*Least Angle Regression*), una versión de *forward step regression*, empezando con  $b_j = 0$  y eligiendo en cada paso el más correlacionado con  $y$  (o  $\hat{y}$ ). Se denomina LARS porque no incluye totalmente al efecto de una variable, sino hasta que no es la más correlacionada con el residuo. En ese caso, entra luego otra y se aumenta en la dirección de MCO, y así sucesivamente. Si un coeficiente alcanza cero, se saca la correspondiente variable. Por lo tanto el objetivo es achicar o disminuir el número de coeficientes a estimar haciendo un *trade off* entre sesgo y varianza (a favor de una menor varianza relativo a MCO) para aumentar de este modo la capacidad predictiva de un modelo. Los estudios de Monte Carlo descritos en Hendry y Doornik (2014)<sup>13</sup>

<sup>12</sup>Para un análisis más detallado ver el capítulo 5 de Sosa Escudero en este libro.

<sup>13</sup>Ver las referencias en el capítulo 17 de Hendry y Doornik (2014).

evalúan *gauge* y *potency* de diferentes enfoques<sup>14</sup>. Los resultados indican que *Autometrics*, comparado con los métodos anteriores, tiene una mejor *performance* en numerosos escenarios, incluso cuando hay más variables que observaciones<sup>15</sup> y, si bien algunos enfoques pueden ser mejores en ciertos casos, en general no serán confiables, ya que en otros casos pueden dar lugar también a resultados extremadamente pobres. Esa variabilidad de *performance* no es observada en *Autometrics*.

Por otra parte, se encuentra disponible un proceso automatizado de búsqueda exhaustiva:

- **GLOBAL SEARCH REGRESSION** (Gluzman y Panigo, 2014) El objetivo de este módulo de STATA es buscar y ordenar aquellos modelos que satisfagan ciertos criterios de selección *in sample* (como bondad del ajuste, dado que los residuos cumplan determinadas propiedades) y, en particular, *out-of-sample* (como por el criterio del error cuadrático medio). Provee también la información (guardada) de los distintos modelos. Debido posiblemente a la diversidad de los criterios, no se conocen estudios específicos de su *performance* en términos de *gauge* y *potency*. Debe notarse, sin embargo, que al ser de búsqueda exhaustiva solo puede aplicarse para los casos de  $K \ll T$  permitiendo como máximo 30 regresores en total.

---

<sup>14</sup>Estos dos conceptos corresponden a las frecuencias obtenidas en los ejercicios de simulación para error de tipo I y potencia (*power*) respectivamente.

<sup>15</sup>Notando que los algoritmos que sean del tipo particular a general pueden usarse para el caso de  $K > T$ .



#### 4. Saturación con *Dummies* y los objetivos de la econometría

El enfoque de Saturación con *Dummies* consiste en la búsqueda de variables *dummies* (*impulse* o *steps*) para cada observación de la muestra. De esta forma, son los datos mismos sin la intervención de la mano del investigador, los que serán informativos sobre la localización de dichas *dummies*.

El trabajo pionero de esta metodología es el de Hendry (1999). En dicho trabajo se trató de identificar a las variables *dummies* (*impulse*) relevantes de un modelo econométrico que se suponía sujeto a valores extremos, siguiendo un simple pero innovativo procedimiento. Primero, incluir la mitad de las variables indicadoras (000...1...0) para cada observación y seleccionar las significativas. Luego, realizar lo mismo con la otra mitad. Posteriormente, combinar las variables *dummies* elegidas y volver a seleccionar. Actualmente, esta selección se ha implementado en *Autometrics* ya que contempla el caso de más variables explicativas que observaciones.

Debemos notar que, si estas variables fuesen redundantes, solo  $\alpha.T$  serán retenidas en promedio por aleatoriedad y por lo tanto, si estamos dispuestos a tolerar la retención de solo un regresor irrelevante, un nivel de significación apropiado será  $\alpha = \frac{1}{T}$ .

La posibilidad de selección de estas variables *dummies* ha sido un aspecto muy importante para el cumplimiento de los tres objetivos básicos de la econometría como explicaremos a continuación.

- Para realizar **inferencia** en los modelos econométricos, la selección de variables indicadoras (*impulse saturation*) para corrección de valores extremos y quiebres conjuntamente con las variables explicativas a incluir en el modelo, nos posibilita contar con estimadores robustos

como Johansen y Nielsen (2009 y 2016) demostraron, tanto para el caso estacionario como de procesos con raíces unitarias<sup>16</sup>.

- Cuando el objetivo es **pronóstico**, el uso de saturación con *dummies* puede ser de gran utilidad. En este caso, tanto el uso de *impulses* (Castle et al., 2012) como *steps* (Castle et al., 2015) proveen una forma de detección de quiebres<sup>17</sup>. Dada la relación entre estos y la capacidad de pronóstico de un modelo<sup>18</sup>, es posible detectar la fecha y duración de los quiebres que servirán para ajustar el modelo (por ejemplo, realizando la corrección de ordenada para llevar el modelo nuevamente (*back-on-track*) o determinar una ventana de estimación apropiada. Debe señalarse que tanto las *impulses* como los *steps* pueden usarse para detectar quiebres y que, a diferencia de otras formas de detección como el test de Bai-Perron, permite su detección tanto al comienzo como al final de la muestra, como puede apreciarse en el análisis de simulación de Castle et al. (2012). Sin embargo, el uso de *step-indicators* es superior a la utilización de la saturación con *impulsos* (y del tradicional test de Chow) cuando ocurren cambios de locación (o sea bajo la usual hipótesis alternativa en estos tests) mejorando así la frecuencia de retención de *dummies* bajo la hipótesis alternativa. También en este caso, el análisis de simulación ha mostrado mejoras con respecto a LASSO (ver Castle et al., 2015). Asimismo, cuando se aplica este tipo de saturación a estudiar sesgos en los errores de pronóstico, se permite la evaluación de sesgos variables a lo largo de los distintos períodos, como veremos posteriormente en la sección 5.

---

<sup>16</sup>Estos autores desarrollaron la teoría asintótica de este caso como la obtención de *estimadores-M*, bajo la hipótesis nula de que no hay valores extremos ni quiebres.

<sup>17</sup>Mientras una variable *dummy* del tipo *impulse* es [000...1...000] una del tipo *step* es [00000...1111111...].

<sup>18</sup>Sobre el uso de los modelos econométricos para el objetivo de pronóstico y su relación con los quiebres ver, entre otros, Clements y Hendry (1998, 1999).

- Para estudiar la exogeneidad de las variables explicativas, la saturación con *dummies* puede ser útil para desarrollar pruebas de *superexogeneidad* automáticas que validan el uso de modelos econométricos condicionales para la **evaluación de políticas**, como una forma de superación de la conocida crítica de Lucas (ver Favero y Hendry, 1992). Superexogeneidad (definida por Hendry et al., 1983) requiere de exogeneidad débil (la validez de estimar solo un modelo condicional) más *invariancia estructural*. Con respecto a esta última propiedad, su evaluación empírica se ha basado en examinar cómo los cambios en los modelos marginales de las variables supuestamente exógenas impactan sobre el modelo condicional de la variable de interés, por ejemplo como propusieron Engle y Hendry, 1993). Si las variables que captan esas no constancias de los modelos marginales (digamos de  $z_t$ ) entran significativamente en el condicional (digamos de  $y_t$  dada  $z_t$ ), entonces  $z_t$  no será superexógena para los parámetros de interés en el modelo condicional. Hendry y Santos (2010) propusieron usar saturación con *dummies* para detectar quiebres y valores extremos en los modelos marginales y evaluar la relevancia de las *dummies* detectadas en el modelo condicional como test de superexogeneidad. Es decir, en una forma muy simplificada, si el modelo condicional de  $y_t$  es

$$y_t = \beta z_t + \mu_t, \quad \mu_t \sim IN(0, \sigma_\mu^2) \quad (4)$$

y el modelo marginal de  $z_t$  es uno autorregresivo que, después de aplicar la saturación por *dummies* resulta,

$$z_t = \pi z_{t-1} + \rho' \mathbf{d}_t + v_t, \quad v_t \sim IN(0, \sigma_v^2) \quad (5)$$

donde el vector  $\mathbf{d}_t$  de variables indicadoras, *impulse* y *steps* fueron seleccionadas para capturar cambios de locación. Entonces, probar la

superexogeneidad  $z_t$  (siendo los parámetros de interés los de modelo condicional de  $y_t$ ) requiere la evaluación conjunta de la significatividad de los efectos de  $\mathbf{d}_t$  en la ecuación (4), es decir  $H_0 : \lambda = 0$  en,

$$y_t = \beta^* z_t + \lambda' \mathbf{d}_t + \eta_t \quad \eta_t \sim IN(0, \sigma_\mu^2) \quad (6)$$

Los resultados de simulación de Hendry y Santos (2010) mostraron que no solo estos tests de superexogeneidad tienen el correcto nivel de significancia, sino también que tienen buena potencia para detectar fallas de superexogeneidad, superando a alternativas anteriormente propuestas para modelar la no constancia de los modelos marginales<sup>19</sup>.

Asimismo, y con relación a la estimación por variables instrumentales para verificar la exogeneidad débil de los regresores, debe señalarse que, al contar con la posibilidad de obtener estimaciones robustas a través de la incorporación de *impulses* y *steps* en la primera etapa, este enfoque nos será de utilidad como una alternativa para fortalecer instrumentos débiles.

En síntesis, los tres objetivos de los modelos econométricos pueden verse beneficiados a partir de la posibilidad de poder aplicar la saturación con *dummies*.

---

<sup>19</sup>Engle y Hendry (1993), por ejemplo, habían propuesto incluir el caso de varianzas no constantes (tipo ARCH) en los modelos marginales, cuya baja potencia para rechazar la hipótesis de constancia fue señalada por Psaradakis y Sola (1996). En cambio, el enfoque de *impulse saturation* aplicado a los modelos marginales detectan apropiadamente cambios de locación que pueden ser utilizados para evaluar su significancia en los modelos condicionales y así verificar superexogeneidad.

## 5. Aplicaciones prácticas

Esta sección presenta algunas aplicaciones realizadas con *Autometrics* con el objetivo de mostrar su versatilidad y así su potenciabilidad futura, ampliando una gran gama de posibilidades en el *practioner corner*.

- **Generalización del estudio de sesgo de pronósticos.** Basado en la propuesta de Mincer y Zarnowitz (1969) por la cual la evaluación de sesgos de los pronósticos pueden realizarse corriendo una regresión de los errores de pronóstico con una constante, Ericsson (2013) extendió dicho enfoque para considerar al caso de sesgos de pronósticos sistemáticos, pero, a su vez, variables en el tiempo. Para ello, sugirió ampliar la ecuación anterior incorporando *impulses* y *steps* obtenidas por Saturación con variables *dummies*. Este enfoque puede ser muy útil para saber en qué períodos aparecen estos sesgos, como en el caso de crisis, y derivar patrones de comportamiento. Por ejemplo, Ahumada y Cornejo (2016) evaluaron de esta forma el papel de los futuros de una *commodity* (soja) para pronosticar los precios *spot*, encontrando sesgos durante la crisis de 2008-2009<sup>20</sup>. Podemos notar también que la evaluación de sesgos en función de distintas variables explicativas se conoce también como tests de *racionalidad fuerte* (Clements y Hendry, 1998, p 57) y en esta línea, *Autometrics* se puede usar para realizar esta evaluación con respecto a amplios conjuntos de información incluyendo numerosos regresores para “explicar” el sesgo.
- **Evaluaciones de no linealidades** La idea de una modelización general a particular incluiría no solo las potenciales variables (con sus rezagos), posibles valores extremos y quiebres, sino también las aproximaciones

---

<sup>20</sup>Ericsson denomina *supersaturation* al caso de la saturación con variables *steps* y generaliza aún más el comportamiento de los sesgos a lo largo del tiempo permitiendo la detección de tendencias quebradas; este último caso lo denomina *ultrasaturation*.

a las no linealidades, por ejemplo por expansiones de Taylor. En particular Castle y Hendry (2010, 2013) consideran que un modelo irrestricto aumentado por un índice de los componentes principales de las potenciales variables explicativas, junto a sus polinomios (de 3er orden, por ejemplo) pueden resultar útiles aproximaciones a conocidos modelos no lineales (como los del tipo de *Smooth Transition Autoregressive (STAR) models*) sugiriendo así una forma *semiautomática* de selección de modelos.

- **Flexibilización de la metodología GVAR.** *Global vector Autoregressions* es la metodología propuesta inicialmente por Pesaran et al. (2004) y extendida por Dees et al. (DdPS, 2007) para modelar las interrelaciones entre los modelos de Corrección de Errores (MCE) correspondientes a un gran número de unidades de corte transversal, inicialmente países o regiones. El objetivo de este enfoque es tener en cuenta importantes variables macroeconómicas de un país o región, tanto las domésticas como las externas, considerando las propiedades de largo plazo (cointegración) y la estructura dinámica de las relaciones. Este análisis se caracteriza por la evaluación de exogeneidad débil (para ver qué países o regiones pueden tomarse como dados para modelar los MCE de cada país o región) y, a partir de allí, realizar una modelación condicional en las variables externas. Los MCEs se obtienen a partir de un modelo VARX inicial, es decir un VAR aumentado con las variables externas a cada país o región, agregadas según el peso en el comercio exterior con los respectivos países, cuyas variables son tomadas como dadas. Esta agregación es lo que permite al enfoque GVAR salvar el problema de la dimensionalidad por las numerosas interrelaciones que surgirían en una estimación irrestricta que llevarían al caso de  $K \gg T$ . Sin embargo, Ericsson (2011) mostró cómo las cuestiones de agregación pueden superarse, y de exogeneidad evaluarse, utilizando *Autometrics* ya que permite considerar el caso de más variables que observaciones.

Un resultado muy interesante de su análisis empírico es que muchas de las restricciones impuestas por GVAR en el caso estudiado por DdPS no son compatibles con los datos cuando se las evalúa a través de las estimaciones realizadas con *Autometrics*. Asimismo, este algoritmo permite la evaluación de constancia de parámetros y, a través de la modelación de los posibles quiebres con variables *dummies*, obtener estimaciones más robustas de las interrelaciones.

- **Evaluación de heterogeneidades en datos de panel y validez del *pooling*.** La selección automática puede servir para determinar apropiadamente los efectos fijos temporales y por grupos, permitiendo una gran cantidad de variables *dummies* aditivas al mismo tiempo que se seleccionan las variables apropiadas del modelo y se tienen en cuenta los efectos de valores extremos. Más aún, debido al esquema de árbol utilizado, permite considerar variables con variación por solo una de las dimensiones (digamos  $z_i$ ) junto a efectos fijos por grupo digamos  $\alpha_i$  y seleccionar entre ellas la mejor opción. También, a través de *dummies* multiplicativas se puede evaluar la constancia de los efectos marginales y elasticidades de las distintas variables explicativas (efectos heterogéneos) y el agrupamiento de distintas unidades de corte transversal en grupos más abarcativos (*pooling*). Esta metodología fue aplicada por primera vez por Ahumada y Cornejo (2015) para el estudio de los precios de un conjunto de *commodities* relevantes para la Argentina. Un resultado importante de este estudio, con relación a este tipo de evaluaciones, es la ausencia de heterogeneidades en el tiempo (antes y después de los 2000 donde generalmente se señala el comienzo del *boom* de *commodities*) pero, en cambio, se detecta un efecto diferencial por *commodity* (de la demanda de China en el precio del petróleo).
- **Estudio de restricciones cuantitativas y regulaciones** Muchos modelos económicos que intentan aproximar decisiones de consumo o producción deben enfrentar el problema de la inclusión de restricciones cuantitativas

(como la prohibición de fumar) y/o regulaciones no asociadas al precio (como las de determinadas especificaciones en los electrodomésticos) que son muy difíciles de incluir en los modelos. Si bien se las ha tratado de captar por la inclusión de variables *dummies*, un problema importante de esta forma convencional es la imposibilidad de estimar el desconocido rezago con que se da su efecto. En cambio, *Autometrics* permite estimarlo a través a través de la saturación con *steps* o, incluso en una forma más parsimoniosa, con las *steps* dadas por el investigador en la fecha de dichas medidas y su rezagos para que los datos indiquen la fecha en que fueron operativas. Una cuestión importante es que la omisión de estas potenciales variables que reflejan restricciones o regulaciones, puede implicar la no detección o la inconsistencia de los efectos de las otras variables relevantes del modelo (como las variables precio o ingreso en una demanda). Un ejemplo de este tipo de estudio se encuentra en Ahumada y Navajas (2016) donde, para el caso de México, se modelaron a través de las descriptas *step dummies*, las medidas de eficiencia energética incorporadas en los electrodomésticos, en un modelo tradicional de las decisiones de consumo de energía en las que se incluyeron conjuntamente el precio de la electricidad, una medida del ingreso y las temperaturas observadas.



## 6. Reflexiones finales

En este capítulo se describieron las principales características y aplicaciones de un algoritmo que automatiza el enfoque conocido como general a particular. Este enfoque tiene como principal rasgo el énfasis en la consideración de las posibles colinealidades, que es en realidad la base del modelo de regresión múltiple y es uno de los temas centrales de la econometría, que enseñamos desde los primeros cursos. Asimismo, la metodología general a particular enfatiza la congruencia del modelo propuesto como forma de aproximación satisfactoria al proceso que genera los datos, en función de la información disponible que posee el investigador. En ese sentido, la selección automática de variables implica un enfoque mucho más completo que la simple búsqueda de un mejor ajuste del modelo econométrico estimado.

*Autometrics* automatiza este proceso de búsqueda de un modelo apropiado a partir de un modelo irrestricto en cuya postulación el econometrista tendrá un papel fundamental. Pero, a partir de allí, la selección de variables por un diagrama de árbol que utiliza el algoritmo, tendrá en cuenta todas las posibles colinealidades siendo, también, eficiente en dicha búsqueda para evitar tiempos tan largos que serían de no relevancia práctica.

Si se toman en cuenta las complejidades de la modelación que requiere, por un lado, tener en cuenta no solo rezagos sino también valores extremos, quiebres y no linealidades y por el otro lado el amplio y creciente conjunto de información que hoy está rápidamente disponible para el investigador, *la maldición de la dimensionalidad* podría ser un obstáculo crítico para la modelación econométrica. Sin embargo, *Autometrics* permite la realización de la búsqueda de variables relevantes también para el caso de  $K > T$ , a través de la consideración de las multicolinealidades por grandes bloques de variables, aunque se debe suponer que  $K \ll T$  en el proceso generador de los datos. La consideración de la posibilidad de valores extremos y quiebres

a través de la saturación con *dummies*, al mismo tiempo que la selección de variables, nos permite detectar efectos que no se podrían detectar, si no se tuviese en cuenta tales *dummies*. Contar con estimaciones robustas nos ayudará con las inferencias que realizaremos con nuestros modelos. Pero a su vez, nos dará nuevos instrumentos para los objetivos de pronóstico y de evaluación de políticas, además de poder realizar muchísimas aplicaciones tanto en el caso de series de tiempo como de panel.

Sin embargo, este algoritmo está actualmente disponible, y sus ventajas comparativas evaluadas, para modelos uniecuacionales encontrándose en evolución la automatización de modelos para sistemas, incluyendo el análisis de cointegración a través de ellos.

En este camino de nuevos desarrollos, una cuestión crítica que permanece es el nivel de significancia (error de tipo I) que el investigador debe definir, en particular para retener variables económicas relevantes y al mismo tiempo, minimizar la de variables *dummies* necesarias para tener en cuenta valores extremos (como en el caso de distribuciones asimétricas y otras formas de no normalidad) y quiebres (como en el caso de la validación del supuesto de homogeneidad de los coeficientes).

Más allá de estas cuestiones que posiblemente serán superadas en un futuro, la liberación del proceso de selección de variables, que puede ser muy tedioso en un mundo que no es ortogonal y en especial, cuando estamos disponiendo de conjuntos de información cada vez más grandes, nos abre la puerta para que los investigadores nos podamos dedicar más plenamente a la creatividad en la modelación econométrica.

## Referencias

- Ahumada, H. (2006). Una nota sobre regresiones con variables integradas. *Ensayos Económicos*, Banco Central de la República Argentina.
- Ahumada, H. (2007). General a Particular, Medio Siglo de Análisis Cuantitativo, en A. Navarro (ed.), *Medio Siglo de Economía*, capítulo 6, Cincuenta Aniversario de la Asociación de Economía Política , Editorial Temas.
- Ahumada, H., y M., Cornejo (2015). Explaining commodity prices by a cointegrated time series-cross section model. *Empirical Economics*, 48(4), 1667-1690.
- Ahumada, H., y M., Cornejo (2016). Out of sample testing price discovery in commodity markets: the case of soybeans. *Agricultural Economics* 47(6), 709-718.
- Ahumada, H., y F., Navajas (2016). Assessing the Impact of Energy Efficiency Standards on Residential Electricity Consumption: the case of Mexico (mimeo), *International Association of Energy Economics (IAEE) Conference*, Bergen, Norway.
- Anderson, T. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *The Annals of Mathematical Statistics*, 33(1), 255-265.
- Campos, J., Ericsson N. y D., Hendry (2005). *General-to-Specific Modelling. The International Library of Critical Writings in Econometrics*, E. Elgar Publishing Ltd, UK.
- Castle, J. L., Doornik J.A. y Hendry, D.F. (2012) Model selection when there are multiple breaks. *Journal of Econometrics*, 169(2), 239-246.
- Castle, J. L., Doornik, J. A., Hendry, D. F., y F., Pretis (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics*, 3(2), 240-264.

- Castle, J.L. y D., Hendry (2010). A low dimension test for non linearity. *Journal of Econometrics*, 158(2), 231-241.
- Castle, J. L. y D., Hendry (2013). Semi-automatic non-linear model selection, en Haldrup, N., Meitz, M., y P., Saikkonen (eds.), *Essays in Nonlinear Time Series Econometrics*, Oxford University Press, Oxford.
- Clements, M. y D., Hendry (1998). *Forecasting economic time series*. Cambridge University Press.
- Clements, M. P. (1999). *Forecasting Non-Stationary Economic Time Series*. Cambridge, MA: MIT Press.
- Dees, S., Mauro, F. D., Pesaran, M. H. y L., Smith (2007). Exploring the international linkages of the euro area: a global VAR analysis. *Journal of applied econometrics*, 22(1), 1-38.
- Doornik, J. A. (2009). Autometrics, en J. L. Castle y N. Shephard (Eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*. Oxford: Oxford University Press.
- Doornik, J. A. y D., Hendry (2015). Statistical model selection with “Big Data”. *Cogent Economics & Finance*, 3(1), DOI: 10.1080/23322039.2015.1045216
- Efron, B., Hastie, T., Johnstone, I. y Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2), 407-499.
- Engle, R. F. y D., Hendry (1993). Testing superexogeneity and invariance in regression models. *Journal of Econometrics*, 56(1-2), 119-139.
- Engle, R. F., Hendry, D. F. y J., Richard (1983). Exogeneity. *Econometrica*, 277-304.
- Ericsson, N. R. (2011). Improving global vector autoregressions. Draft, *Board of Governors of the Federal Reserve System*, Washington, DC, June.
- Ericsson, N. R. (2013). How Biased Are US Government Forecasts of the Federal Debt? Washington, DC: Board of Governors of the Federal Reserve System.

- Ericsson, N. R. y J., MacKinnon (2002). Distributions of error correction tests for cointegration. *The Econometrics Journal*, 5(2), 285-318.
- Favero, C. y Hendry, F. D. (1992). Testing the Lucas critique: A review. *Econometric Reviews*, 11(3), 265-306.
- Favero, C. A. (2001). *Applied macroeconometrics*. Oxford University Press on Demand.
- Gilbert, C. L. (1986). Practitioners' corner: Professor Hendry's Econometric Methodology. *Oxford Bulletin of Economics and Statistics*, 48 (3), 283-307.
- Gluzman, P. y D., Panigo (2014). Global Search Regression (GSREG): A new automatic model selection technique for cross section, time series and panel data regressions. Asociación Argentina de Economía Política, XLIX Reunión Anual, Posadas.
- Hendry, D. F. (1995). *Dynamic econometrics*. Oxford University Press.
- Hendry, D. F. (1999). An econometric analysis of US food expenditure, 1931-1989, en J.R. Magnus y M.S. Morgan (eds.), *Methodology and Tacit Knowledge: Two Experiments in Econometrics*, pp. 341-61. Chichester: John Wiley and Sons.
- Hendry, D. F. (2006). Robustifying forecasts from equilibrium-correction systems. *Journal of Econometrics*, 135, 399-426.
- Hendry, David F. y A., Doornik (2014). *Empirical model discovery and theory evaluation: automatic selection methods in econometrics*. MIT Press.
- Hendry, D. F., Johansen, S., y C., Santos (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, 23, 317-335.
- Hendry, D. F. y H., Krolzig (2004). We ran one regression. *Oxford bulletin of Economics and Statistics*, 66(5), 799-810.
- Hendry D.F. y Santos C (2010). An automatic test for super exogeneity, en Watson, MW, Bollerslev T, y J., Russell (eds), *Time series econometrics: essays in Honor of Robert F.* Oxford University Press, Oxford.

- Hoover, K. D. y S., Perez (1999). Data mining reconsidered: encompassing and the general to specific approach to specification search. *The econometrics journal*, 2(2), 167-191.
- Johansen, S. y B., Nielsen (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, 43(2), 321-348.
- Johansen S. y Nielsen B (2009). An analysis of the indicator saturation estimator as a robust regression estimator, en Castle J y N. Shepard (eds), *The methodology and practice of econometrics: a Festschrift in Honour of David F. Oxford University Press*, Hendry, Oxford.
- Lovell, M.C. (1983). Data Mining. *Review of Economics and Statistics*, 65,1-12.
- Mincer, J. A. y Zarnowitz, V. (1969). The evaluation of economic forecasts, en *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, 1-46, NBER books.
- Psaradakis, Z. y Sola, M. (1996). On the power of tests for superexogeneity and structural invariance. *Journal of Econometrics*, 72(1), 151-175.
- Perez Amaral, T., Gallo, G. M. y White, H. (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics*, 65, 821-838.
- Pesaran, M. H., Schuermann, T. y Weiner S. (2004). Modeling Regional Interdependencies Using a Global Error correcting Macroeconometric Model, *Journal of Business and Economic Statistics*, 22(2), 129-181.
- Sims, C. A., Stock, J. H. y Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica: Journal of the Econometric Society*, 113-144.
- Sucarrat, G., Pretis, F. y J., Reade (2017). General-to-Specific (GETS) Modelling and Indicator Saturation, R Package Version 0.12.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.





## **Capítulo 3. Análisis estructural de datos de subastas**

M. Florencia Gabrielli

### **1. Introducción**

Las subastas son mecanismos de mercado que se utilizan ampliamente para asignar contratos públicos, activos financieros, productos agrícolas, recursos naturales, obras de arte y electricidad, entre otros. También el desarrollo de subastas por internet ha tomado gran relevancia en los últimos años.

Las subastas tienen reglas bien definidas que pueden plasmarse en un modelo económico. El concepto de Equilibrio Bayesiano de Nash (BNE, *Bayes Nash Equilibrium*) ha permitido la modelización de las subastas en donde un vendedor enfrenta un número limitado de participantes que se comportan estratégicamente. Consideraremos en la mayor parte de este capítulo el caso en que un vendedor ofrece un ítem a un conjunto de compradores o participantes. De manera simétrica se podría pensar que un comprador está intentando adquirir un bien o servicio y ejecuta una subasta que involucra un conjunto

de vendedores, cada uno de los cuales puede proveer dicho bien. Este tipo de “licitación” o subasta es típicamente utilizada por los gobiernos para comprar bienes. En cualquier caso, las subastas son, desde una perspectiva teórica, juegos de información incompleta donde las asimetrías de información entre el subastador (vendedor/comprador) y los participantes (quienes quieren adquirir o proveer el bien) y entre los mismos participantes juegan un rol crucial<sup>1</sup>.

El supuesto implícito que se hace al modelar subastas es que cada jugador posee un valor intrínseco por el bien que se está subastando; es decir el jugador está dispuesto a comprar dicho bien por un precio que no va a superar en ningún caso ese valor intrínseco. Nos referimos a ese valor como la *valuación* o *valoración verdadera* que el jugador tiene por el bien.

En el caso en que se venda un único bien (indivisible), podemos distinguir cuatro tipos principales de subastas (y muchas variantes de estos cuatro tipos).

1. Subastas ascendentes o Inglesas. Este tipo de subastas se ejecutan en tiempo real y los participantes están físicamente (o electrónicamente) presente. El vendedor va incrementando gradualmente el precio (desde una base prefijada), y los compradores van saliendo del juego hasta que finalmente queda uno quien es el que gana al precio del momento. Las subastas orales en las cuales la gente hace ofertas cantadas, o hace ofertas de manera electrónica, son variantes de este tipo de subasta ascendente.
2. Subastas descendentes u Holandesas. También son subastas interactivas que se llevan a cabo en tiempo real pero en este caso el vendedor comienza bajando el precio a partir de un precio inicial alto hasta el momento en que algún participante acepta la oferta y paga el precio actual. Este formato de subasta se llama también Holandesa porque las flores en los Países Bajos se han vendido de esta manera por mucho tiempo.

---

<sup>1</sup>Nos referimos con el término participantes a los “jugadores”, es decir quienes compiten por obtener o proveer el bien según sea una subasta o un licitación, respectivamente.

3. Subastas de primer precio a sobre cerrado. En este tipo de subasta, los participantes le entregan al vendedor de manera simultánea “ofertas selladas”. La terminología proviene del formato original que tenían este tipo de subastas, en las cuales las ofertas eran escritas en un papel y se entregaban en sobre sellado al vendedor quien abría los sobres al mismo tiempo para determinar el ganador. El participante que hizo la oferta más alta es el ganador y el pago que realiza por el bien es igual a dicha oferta.
4. Subastas de segundo precio a sobre cerrado. También llamadas subastas de Vickrey. Los participantes entregan simultáneamente ofertas selladas al vendedor; el ganador es quien hace la mayor oferta pero el pago que efectúa es igual al valor de la segunda mejor oferta (la segunda más alta). El nombre de subasta de Vickrey es en honor a William Vickrey, premio Nobel en Economía en 1966, quien fue el primero en escribir sobre subastas desde una perspectiva de teoría de juegos, incluyendo este tipo de subasta.

En este capítulo nos concentramos en subastas de primer precio a sobre cerrado. Como mencionamos anteriormente, el análisis de una licitación donde el ganador es quien hace la menor oferta y recibe como pago dicho valor, es totalmente simétrico teniendo en cuenta los cambios de signo correspondientes. Analizar este tipo de mecanismo no es restrictivo. La mayoría de las bases de datos de subastas disponibles para investigación empírica provienen de este formato. Más aún, otros formatos son estratégicamente equivalentes o posibles de ser analizados desde una perspectiva similar.

Hasta no hace mucho el análisis empírico de las subastas se limitaba a testear algunas predicciones de los modelos de teoría de juegos subyacentes (ver Porter, 1995) para una recopilación exhaustiva del Enfoque de Forma Reducida. A esto se lo conoce como Análisis de Forma Reducida (*Reduced Form Approach*). Este enfoque no permite realizar evaluaciones de política que requieren el conocimiento de la estructura informacional del juego como por

ejemplo la elección de un precio de reserva y el mecanismo de subasta que generaría una mayor recaudación para el vendedor/comprador.

El Enfoque Estructural (*Structural Approach*) permite abordar tales cuestiones asumiendo que las ofertas observadas son las ofertas de equilibrio de algún modelo de subasta<sup>2</sup>. Específicamente este enfoque explora la relación de equilibrio  $b_i = s_i(v_i)$ , donde  $b_i$  denota la oferta (observada) realizada por el jugador  $i$ ,  $s(\cdot)$  es la función que caracteriza a la estrategia de equilibrio y  $v_i$  es la valoración privada (no observada) que el jugador tiene por el bien. A los efectos de modelar comportamiento, se asume que la información privada de los jugadores (i.e., sus valuaciones) proviene de alguna distribución que es de conocimiento común para todos los jugadores, aunque cada jugador solamente conoce su propia valoración y no la de sus contrincantes. Esta distribución y las preferencias de los individuos (caracterizadas por una función de utilidad) son los elementos clave que explican el comportamiento, es decir forman la estructura del modelo. En otras palabras son elementos estructurales del modelo econométrico inducido por las ofertas observadas (la data). Este capítulo tiene por objetivo describir los aspectos centrales y algunas de las contribuciones más significativas sobre el Enfoque Estructural para analizar datos de subastas. Los aportes que se han desarrollado en los últimos 20 años son numerosos y sería imposible citarlos a todos.

Una dificultad importante asociada al Enfoque Estructural radica en la complejidad numérica o en la forma implícita asociada a las estrategias de equilibrio. Hay tres aspectos importantes que deben considerarse.

En primer lugar, el tema de la *Identificación* es decir si la estructura de la subasta puede ser recuperada en forma unívoca a partir de las variables observables (las ofertas) minimizando las restricciones paramétricas. En otras

---

<sup>2</sup>El tipo de supuestos que se hace sobre las valuaciones (por ejemplo, si son independientes o no, si existe un componente común en la estructura de información o no existe tal componente) determina el paradigma que se utiliza. En la próxima sección detallamos los distintos modelos o paradigmas que hay en la literatura de subastas.

palabras la identificación implica establecer si los modelos correspondientes pueden ser discriminados entre sí a partir de las variables observables. Una segunda cuestión importante concierne a la validación del modelo, es decir si el modelo (teórico) impone restricciones testeables sobre los datos. Finalmente, una tercera dificultad está asociada al desarrollo de métodos de estimación aplicables.

En su trabajo seminal sobre el Análisis Estructural de las Subastas, Laffont y Vuong (1996) destacan que hay razones prácticas y teóricas que han llevado a desarrollar el área de “Econometría de Subastas”. La amplia utilización de estos mecanismos para asignar bienes y servicios ha generado importantes bases de datos que pueden utilizarse para analizar estos mercados. Por otro lado, se han hecho contribuciones significativas a la teoría de Organización Industrial y ahora el desafío radica en llevar dichas contribuciones a la práctica. El trabajo de Laffont y Vuong (1996) establece un marco unificado para analizar modelos de subastas desde el Enfoque Estructural.

El resto de este capítulo se organiza de la siguiente manera. La sección 2 contiene el modelo canónico a partir del cual se puede establecer un marco de referencia unificado para analizar subastas. Dada la estrecha relación entre el modelo teórico y el correspondiente modelo econométrico que prevalece en el enfoque estructural, la sección 3 presenta la caracterización, identificación y restricciones testeables de los distintos modelos teóricos de subastas. En este sentido se presenta el Paradigma de Valuaciones Afiliadas que distingue dos casos polares, las Valuaciones Privadas y las Valuaciones Comunes. A su vez, cada modelo puede clasificarse en simétrico o asimétrico. Cada uno de estos casos conduce a distintos modelos de subastas. La sección 4 describe los métodos econométricos desarrollados desde el enfoque estructural para estimar los distintos modelos dentro del paradigma de valuaciones privadas. En la sección 5 se describe en detalle el método no paramétrico en dos etapas de Guerre et al. (2000), quienes introdujeron el uso de procedimientos indirectos

no paramétricos para estimar modelos de subastas, y se menciona brevemente una extensión semi-paramétrica del mismo. La sección 6 muestra cómo incorporar comportamientos no cooperativos (por ejemplo, colusión) en los modelos y cómo pueden adaptarse las estrategias econométricas. Finalmente la sección 7 recoge las principales conclusiones.

## 2. Modelo canónico

El marco teórico adoptado es el modelo de Wilson (1977), en donde hay un objeto indivisible que es subastado a  $n$  oferentes. La función de utilidad individual es  $U_i = U(v_i, c)$  que es creciente en cada uno de sus argumentos.  $v_i$  representa la información privada de cada jugador y  $c$  es el componente común. Para entender el rol del componente común, es útil pensar en subastas en donde el bien que se negocia será luego re-venido en un mercado secundario. En estos casos, es natural pensar que el valor que los distintos participantes le asignan a dicho bien estará altamente influenciado por el valor de re-venta del mismo, que en términos del modelo lo representamos a través del valor común. Por lo tanto, las valuaciones privadas se deben entender como las “señales” o información privada que tienen los distintos participantes sobre dicho valor. Si bien el valor  $c$  es único (solo va a existir un único precio de re-venta) los individuos pueden tener información privada (ruidosa) sobre el mismo.

El vector  $(v_1, \dots, v_n, c)$  se distribuye en forma conjunta de acuerdo a la función  $F(\cdot)$ . Se asume que cada jugador solamente sabe su propia señal o valoración,  $v_i$  (no conoce las otras señales ni el componente común), es decir el juego es de información incompleta, Harsanyi (1967). Finalmente se asume que  $n$ ,  $U(\cdot)$  y  $F(\cdot)$  son de conocimiento común. Este es el modelo general, AV (*Affiliated Value Model*). Intuitivamente decir que  $F(\cdot)$  es *afiliada* implica

que una valoración alta de un jugador hace que los otros jugadores tengan valoraciones también más altas con una mayor probabilidad. En este sentido, el concepto de afiliación se corresponde con una noción fuerte de correlación positiva.

Si  $F(., \dots, .)$  es simétrica en  $v_i$ , entonces el modelo es simétrico, de otro modo el modelo es asimétrico. En un modelo simétrico, todos los jugadores son ex-ante iguales, por lo tanto las estrategias de equilibrio son iguales. Esto significa que es suficiente analizar el comportamiento de uno de los jugadores, lo que se reduce a obtener una ecuación o condición de primer orden. En contraposición, en un modelo asimétrico, los jugadores son ex-ante diferentes, podríamos tener tantos “tipos de jugadores” como cantidad de jugadores. También es posible que existan menos tipos de jugadores que cantidad de jugadores si agrupamos en distintos subconjuntos a los jugadores del mismo tipo, en cuyo caso se considera que dentro de un mismo tipo los jugadores son simétricos. En términos de la solución de estos modelos, la misma se caracteriza a través de un sistema de ecuaciones de primer orden donde la cantidad de ecuaciones involucrada es igual a la cantidad de tipos de jugadores en cuestión.

Una estrategia para un jugador  $i$  es una función  $s(v) = b$  entre su verdadera valoración y una oferta,  $b$ , no negativa. Hacemos los siguientes supuestos (estándares en la literatura) sobre estas estrategias<sup>3</sup>.

- (i)  $s(\cdot)$  es una función estrictamente creciente y continuamente diferenciable; de manera que, en particular, si dos participantes tienen distintas valoraciones por el objeto, entonces cada uno de ellos presentará distintas ofertas.
- (ii)  $s(v) \leq v$  para todo  $v \in [\underline{v}, \bar{v}]$ , es decir los jugadores pueden ofertar menos que su valoración, pero nunca van a ofertar un valor mayor a la misma. Dado que las ofertas son no-negativas siempre, este supuesto

<sup>3</sup>Para simplificar la notación nos abstraemos de posibles asimetrías entre jugadores. En juegos asimétricos hay que especificar estos supuestos para cada tipo de jugador.

implica que  $s(\underline{y}) = \underline{y}$ . A esta última condición se la suele denominar condición de “contorno” o “frontera” ya que impone una restricción al comportamiento del jugador que tiene la menor valuación, es decir el jugador que se encuentra en la frontera del soporte de las valuaciones.

Para caracterizar el equilibrio bayesiano de Nash (BNE) hay que destacar que cada jugador maximiza su pago esperado, esto es, la ganancia de la subasta multiplicada por la probabilidad de ganar. La estrategia de equilibrio “ $b$ ” se define como un BNE donde cada jugador escoge su mejor oferta (o respuesta) dada su información privada (señal o valuación), dadas las estrategias de los otros participantes y sus expectativas racionales sobre las señales de los otros jugadores. Matemáticamente, estas estrategias de equilibrio son las soluciones de un sistema de ecuaciones diferenciales<sup>4</sup>.

Es común que en las subastas exista lo que se denomina un “precio de reserva”, denotado por  $p_0$ , que es el precio por debajo del cual el vendedor no está dispuesto a vender el bien. Para simplificar vamos a suponer que este precio de reserva es no vinculante (*non binding*), esto es  $p_0 = \underline{y}$ . Este supuesto dice que todos aquellos que quieren entrar a la subasta lo hacen, es decir, el precio de reserva no actúa como un mecanismo de exclusión.

### 3. Modelos de subastas, identificación y restricciones teóricas

La hipótesis que se mantiene en el Enfoque Estructural es que las ofertas observadas son las ofertas de equilibrio del modelo de subasta bajo análisis. Dado  $n$  y una estructura  $[U(\cdot), F(\cdot)]$ , la estrategia de equilibrio del juego induce al siguiente modelo econométrico,

---

<sup>4</sup>Como se verá más adelante en algunos modelos la solución no tiene forma analítica o cerrada.



$$b_i = s_i(v_i, n, U, F).$$

Es importante advertir que  $v_i$  es una variable aleatoria no observable, por lo tanto las ofertas  $b_i$  son también aleatorias y se distribuyen de acuerdo a la función  $G(\cdot)$ , digamos, que está determinada por los elementos estructurales del modelo. En este sentido, la función  $G(\cdot)$  depende de  $F(\cdot)$  a través de dos canales.

1. A través de los  $v_i$ 's, ya que los mismos se distribuyen de acuerdo a  $F(\cdot)$ .
2. A través de la estrategia de equilibrio,  $s(\cdot)$  (ver la ecuación (1) más abajo por ejemplo).

Estas características propias de los modelos de subastas complican la identificación y caracterización de restricciones teóricas. Es por esto que para establecer identificación se analizarán modelos en el marco de  $n$  dado y ofertas observables dadas, es decir que esta es la única información disponible para el econométrista para identificar el modelo, esto es para recuperar las valuaciones (que caracterizan a la estructura del modelo) a partir de las observaciones (las ofertas).

Es necesario precisar lo que se entiende por modelos que no pueden identificarse o distinguirse uno de otro. La distribución  $G(\cdot)$  de las ofertas  $(b_1, \dots, b_n)$  es *racionalizada* por una estructura  $[U(\cdot), F(\cdot)]$  si  $G(\cdot)$  es la distribución de equilibrio del juego correspondiente. Damos al respecto la siguiente definición.

**Definición:** dos modelos  $M$  y  $\tilde{M}$  son observacionalmente equivalentes dadas las ofertas (observadas)  $(b_1, \dots, b_n)$  si ambos racionalizan la misma distribución de ofertas de equilibrio, esto es,  $G(\cdot) = \tilde{G}(\cdot)$

Es decir, si dos modelos distintos nos proveen la misma función de distribución de las ofertas de equilibrio, será imposible, a partir de dicha función,

discriminar entre los dos modelos. Por ejemplo un modelo en donde las valuaciones son afiliadas y hay componente común no puede distinguirse, a partir de las ofertas, de un modelo en donde las valuaciones son afiliadas y no existe componente común como se discute más abajo.

A continuación presentamos los distintos modelos de subastas comenzando por el modelo con valuaciones afiliadas. Este modelo es el que impone la menor estructura (o supuestos) sobre las valuaciones y el componente común, lo cual lo convierte en un modelo muy general. No obstante, a los efectos de identificar el modelo, veremos que es importante sacrificar generalidad para lograr identificación, lo que nos lleva a modelos con una mayor cantidad de supuestos sobre la naturaleza de las valuaciones y el componente común. En cada caso veremos la versión simétrica y asimétrica del juego correspondiente.

### 3.1. El modelo AV simétrico

#### 3.1.1. Caracterización

El modelo AV (*Affiliated Value Model*) simétrico ha sido caracterizado por Wilson (1977) y Milgrom y Weber (1982). De estos trabajos sabemos que el equilibrio de Nash de este juego es una función,  $s(v)$ , estrictamente creciente en  $v$  y surge del siguiente problema de optimización.

$$\text{máx } E[(U(v_i, c)) - b_i \mathbf{I}(B_i \leq b_1) | v_i]$$

con  $B_i = s(y_i)$ ,  $y_i = \text{máx}_{j \neq i} v_j$ ,  $s(\cdot)$  es la estrategia de equilibrio. Se puede demostrar que la solución está dada por,

$$b_i = s(v_i) = V(v_i, v_i) - \int_{\underline{v}}^{v_i} L(\alpha|v_i) dV(\alpha, \alpha) \quad (1)$$

donde  $L(\alpha|v_i) = \exp[-\int_{\alpha}^{v_i} f_{y_1|v_1}(s|s)/F_{y_1|v_1}(s|s)ds]$  y  $V(v_i, y_i) = E[U(v_i, c)|v_i, y_i]$ , la condición de contorno está dada por  $s(\underline{v}) = V(\underline{v}, \underline{v})$ .

Como se puede observar en este caso tenemos solución analítica.

Hay dos casos especiales que han sido intensamente estudiados. En primer lugar el modelo simétrico con valuaciones privadas independientes, modelo IPV (*Independent Private Value*) en donde se asume que  $U_i = v_i$  y  $v_i$  es independiente e idénticamente distribuida (iid) como  $F(\cdot)$  (ver Vickrey, 1961, Riley y Samuelson, 1981). En segundo lugar el modelo simétrico de valuaciones comunes, el modelo CV (*Common Value Model*) en donde la función de utilidad es  $U_i = c$ , con  $v_i$  iid dado  $c$  como  $F_{v|c}(\cdot|\cdot)$ . En cualquiera de estos dos casos, la solución de equilibrio (1) se simplifica (ver Rothkopf, 1969, y Wilson, 1969).

### 3.1.2. Identificación

Como mencionamos más arriba, el paso previo para poder estimar un modelo de subastas es establecer la identificación de los elementos estructurales a partir de los datos disponibles. Dado que los elementos estructurales en los modelos de subastas son la función de distribución conjunta de las valuaciones y el componente común,  $F(\cdot)$ , y la función de utilidad,  $U(\cdot)$ , que también depende de  $v$  y  $c$ , la identificación del modelo implica establecer si es posible obtener las valuaciones y el componente común (que son variables no observables) a partir de las ofertas observadas (y de la cantidad de jugadores) utilizando

las condiciones de primer orden que son las que naturalmente establecen una relación entre dichas variables<sup>5</sup>.

A partir del trabajo seminal de Laffont y Vuong (1996) tenemos una serie de resultados que establecen en qué casos puede lograrse la identificación de los distintos modelos y en qué casos los modelos imponen restricciones testeables con el objetivo de poder validarlos.

Laffont y Vuong (1996) comienzan analizando la identificación del modelo más general posible, el modelo AV y luego siguen con modelos más específicos, es decir modelos con más restricciones. Naturalmente la identificación se simplifica si el modelo es menos general.

En este sentido el primer resultado (Proposición 1) que tenemos es de alguna manera negativo y establece que no puede distinguirse a partir de la cantidad de ofertas solamente un modelo AV simétrico de un modelo APV (*Affiliated Private Value*) simétrico. Es decir los modelos AV y APV son observacionalmente equivalentes. La prueba de este resultado se basa en la siguiente observación.  $U_i = U(v_i, c)$  puede reemplazarse por  $\tilde{U}_i = \tilde{v}_i$  con  $\tilde{v}_i = V(v_i, v_i)$ . Es decir, la dependencia entre utilidades dada por el componente común,  $c$ , puede ser reemplazada por dependencia (en forma apropiada) entre señales privadas. Por lo tanto, dado este resultado es que analizaremos la identificación del modelo APV en donde  $U_i = v_i$  sin pérdida de generalidad, ya que este modelo es el más general que puede identificarse.

Dado el resultado anterior, el siguiente resultado es naturalmente el que establece que el modelo APV con jugadores simétricos está identificado. Notemos que en este paradigma las valuaciones se asumen afiliadas pero privadas, es decir puede estar correlacionadas pero no existe componente común, por lo tanto la naturaleza de la correlación no se puede deber a la existencia de un componente común. La identificación se logra básicamente

---

<sup>5</sup>Como veremos más adelante, no siempre se logra identificar al componente común,  $c$  a partir de las ofertas solamente.

debido a que las valuaciones (no observadas) y las ofertas (observadas) están relacionadas a través de una estrategia de equilibrio estrictamente creciente con lo cual  $F(\cdot)$  puede identificarse a partir de  $G(\cdot)$ , o lo que es lo mismo las valuaciones se pueden identificar a partir de las ofertas. Un caso particular intensamente estudiado en la literatura es el que asume que las valuaciones privadas son independientes (se supone que no hay afiliación) y se llega de esta manera al paradigma IPV. Como es de esperar este modelo está identificado. El trabajo de Guerre et al. (2000) contiene la prueba formal de este resultado. La intuición detrás de este resultado es similar a la de la proposición que establece la identificación de modelos APV mencionada más arriba.

Cuando se cambia del paradigma de valuaciones privadas al de valuaciones comunes, los resultados no son positivos en cuanto a identificación. En particular los modelos CV simétricos generalmente carecen de identificación. Este resultado se desprende del hecho de que modelos con afiliación (AV) y modelos con afiliación y valores privados (APV) no se pueden distinguir (Proposición 1 de Laffont y Vuong, 1996) y por lo tanto como el modelo CV es un caso particular de un modelo AV, un modelo CV en general no puede distinguirse de un modelo APV. La intuición de este resultado puede establecerse observando que si se definen nuevas señales,  $\tilde{v}_i$  como transformaciones estrictamente crecientes de las señales  $v_i$  originales, se puede ver que la identificación del modelo no es posible. (Recordar que se ha definido al modelo CV con  $U_i = c$  y  $v_i$  iid dado  $c$ ).

### 3.1.3. Restricciones testeables

Una vez que se ha logrado identificar a algún modelo, el siguiente paso es establecer si dicho modelo teórico impone restricciones testeables que

sirvan para validarlo a partir de los datos. En este sentido Laffont y Vuong (1996) proporcionan una serie de resultados. Para establecer estas restricciones sobre los datos mantenemos la hipótesis de que  $F(\cdot)$  es simétrica y afiliada. Denotemos por  $b_i = s_0(v_i)$ , a la función que verdaderamente utilizan los jugadores para armar sus estrategias (notar que  $s_0$  puede ser distinta de  $s$ , que es el BNE) para todo  $i$ , se sigue que  $G(\cdot)$  es simétrica y afiliada en  $[\underline{b}, \bar{b}]^n$ .

La Proposición 1 de Laffont y Vuong (1996) analizada más arriba implica que las restricciones que se imponen sobre un modelo AV son las mismas que pueden imponerse sobre un modelo APV (más allá de la simetría y la afiliación). Definimos la siguiente función que necesitaremos en el próximo resultado. Sea  $G_{B_1|b_1}(\cdot|\cdot)$  la distribución condicional de  $B_1 = \max_{j \neq 1} b_j$  dado  $b_1$ , es decir es la distribución de la máxima oferta rival. Laffont y Vuong (1996) establecen que la distribución  $G(\cdot)$  puede racionalizarse a través de un modelo APV si y solo si la función

$$\xi(b) = b + \frac{G_{B_1|b_1}(b|b)}{g_{B_1|b_1}(b|b)}$$

es estrictamente creciente en  $b \in [\underline{b}, \bar{b}]^n$ .

Hay que notar que  $s_0(\cdot)$  (la estrategia que utilizan los participantes) no es necesariamente  $s(\cdot)$  (la estrategia de equilibrio) que se deduce de las condiciones de primer orden (CPO) del juego. Por lo tanto para testear si los jugadores se comportan de acuerdo a lo que predice la teoría de juegos hay que corroborar que  $\xi(b)$  sea estrictamente creciente, caso contrario  $s_0(\cdot)$  no puede ser un equilibrio bayesiano de Nash. Un resultado existe para los modelos IPV, ya que estos modelos constituyen un caso particular de los modelos APV. El trabajo de Guerre et al. (2000) contiene un argumento formal que prueba esta proposición. Más aún al ser los modelos IPV más restrictivos, por el hecho de suponer independencia de las valuaciones, se tiene adicionalmente otro test

---

<sup>6</sup>Recordar que, la distribución  $G(\cdot)$  de las ofertas  $(b_1, \dots, b_n)$  es *racionalizada* por una estructura  $[U(\cdot), F(\cdot)]$  si  $G(\cdot)$  es la distribución de equilibrio del juego correspondiente.

sobre la naturaleza de las funciones que racionalizan a las ofertas. En particular en este paradigma con valuaciones privadas independientes, la distribución de las ofertas  $G(\cdot)$  puede racionalizarse si y solo si se cumple lo siguiente.

1.  $G(b_1, \dots, b_n) = \prod_{i=1}^n G_0(b_i)$  para alguna distribución  $G_0(\cdot)$  con soporte en  $[b, \bar{b}]$ .

2. La función

$$\xi(b) = b + \frac{1}{n-1} \frac{G_0(b)}{g_0(b)}$$

es estrictamente creciente en  $b \in [b, \bar{b}]$ .

La primer condición establece que las ofertas son independientes e idénticamente distribuidas (iid). La segunda es igual a la restricción en modelos APV. Notar que en modelos IPV la función de distribución de las ofertas relevantes es simplemente la distribución conjunta.

En cuanto a los modelos de CV, las restricciones que impone la teoría de juegos no están tan desarrolladas. Algunos resultados recientes sobre el tema pueden consultarse en Hendricks et al. (2003) y Haile et al. (2003).

Es importante aclarar que dado que las ofertas de equilibrio en un modelo CV deben ser dependientes mientras que un modelo IPV deben ser independientes, los modelos IPV y CV en principio pueden distinguirse.

### 3.2. El modelo AV asimétrico

La teoría general está menos desarrollada para estos modelos. Las asimetrías pueden darse debido a que algunos jugadores estén mejor informados que otros, también debido a que algunos jugadores pueden organizarse mejor a través de consorcios (cárteles), otra fuente de asimetrías son las distintas características

de los participantes como el tamaño y la localización. Consideraremos dos casos especiales.

1. Modelo asimétrico IPV
2. Modelo asimétrico CV

### 3.2.1. Caracterización de los modelos bajo análisis

1. Modelo asimétrico IPV

En este caso tenemos  $U_i = v_i$ , y consideraremos  $v_1 \sim F_1(\cdot)$ ,  $v_j \sim F_2(\cdot)$ ,  $j \neq 1$ ,  $F_1(\cdot) \neq F_2(\cdot)$ . Es decir analizamos dos tipos de jugadores solamente. Es evidente que el problema del jugador 1 es diferente al de los demás. Las estrategias de equilibrio  $s_1(\cdot)$  y  $s_2(\cdot)$  por lo general no tienen solución analítica. Hay que recurrir a métodos numéricos.

2. Modelo asimétrico CV

En este modelo suponemos que un jugador está plenamente informado y el resto está desinformado.  $U_i = c$  para todos los jugadores con  $v_1 = c$  y  $v_j = \phi$  para  $j \neq 1$ . También se tiene que  $c \sim F_c(\cdot)$ . La estrategia de equilibrio del jugador informado es,

$$b_1 = s_1(v_1) = v_1 - \int_{\underline{v}}^{v_1} \frac{F_c(u)}{F_c(v_1)} du. \quad (2)$$

Los jugadores no informados deben adoptar estrategias mixtas ( $M(\cdot)$ ) en el intervalo  $[\underline{v}, E(c)]$  de forma que su oferta máxima simule la distribución del jugador informado, esto es,

$$M^{n-1}(\cdot) = F_c[s^{-1}(\cdot)].$$



### 3.2.2. Identificación y restricciones testeables

Para los dos casos caracterizados trataremos de responder los siguientes interrogantes. ¿Se puede identificar cada modelo?, ¿hay restricciones testeables que cada modelo imponga sobre la distribución de las ofertas?, ¿se pueden distinguir estos modelos asimétricos a partir de las ofertas observadas? y finalmente, ¿se puede distinguir un modelo asimétrico de uno simétrico?. Como es de esperar el resultado de identificación para el modelo IPV asimétrico es positivo, aún cuando en este modelo el equilibrio bayesiano de Nash no tiene una solución analítica. Más aún, la distribución  $G(\cdot)$  puede racionalizarse si y solo si se cumplen las siguientes condiciones

1.  $G(b_1, \dots, b_n) = G_1(b_1) \prod_{i=2}^n G_2(b_i)$  para alguna distribución  $G_1(\cdot)$  y  $G_2(\cdot)$ .
2. Las funciones

$$\xi_1(b) = b + \frac{1}{n-1} \frac{G_2(b)}{g_2(b)},$$

$$\xi_2(b) = b + \frac{1}{\frac{g_1(b)}{G_1(b)} + (n-2) \frac{g_2(b)}{G_2(b)}}.$$

son estrictamente crecientes en  $b \in [\underline{b}, \bar{b}]$ .

Lo que es más sorprendente, probablemente, es que también se consigue identificación para el modelo CV caracterizado más arriba. En este caso las restricciones teóricas sobre el mismo son una adaptación natural de las restricciones que hemos venido analizando pero para el presente modelo. Es decir, la distribución  $G(\cdot)$  puede racionalizarse si y solo si se cumple,

1.  $G(b_1, \dots, b_n) = G_1(b_1) \prod_{i=2}^n M(b_i)$  para alguna distribución  $G_1(\cdot)$  y  $M(\cdot)$ .
2.  $G_1(\cdot) = M^{n-1}(\cdot)$ , y la función

$$\xi(b) = b + \frac{1}{(n-1)} \frac{M(b)}{m(b)}$$

es estrictamente creciente en  $b \in [\underline{b}, \bar{b}]$ .

A pesar de que cada modelo caracterizado está identificado, nos podemos preguntar si estos dos modelos asimétricos pueden ser discriminados a partir de las ofertas observadas. Lamentablemente no tenemos un resultado positivo al respecto. En su Proposición 9, Laffont y Vuong (1996) establecen que cualquier modelo CV asimétrico es observacionalmente equivalente a algún modelo IPV asimétrico. Finalmente, para responder el último interrogante planteado es suficiente con considerar el caso  $n = 2$ . Luego las condiciones testeables sobre un modelo CV asimétrico y las que se desprenden de uno simétrico son equivalentes, lo que lleva a concluir que cualquier modelo CV asimétrico no necesariamente se distingue de un modelo simétrico.

#### 4. Estimación estructural

En esta sección estudiaremos las contribuciones econométricas para el análisis estructural de subastas de primer precio (FPA, *First Price Auctions*) en el marco de modelos con valoraciones privadas tanto en su versión simétrica como asimétrica. El hecho de centrar nuestra atención en modelos de subastas de primer precio no es una limitante ya que los métodos que discutiremos pueden aplicarse también a subastas descendentes porque en ambos casos las estrategias de equilibrio son equivalentes. Más aún, en la práctica las subastas de primer precio constituyen el formato más comúnmente utilizado. En los mecanismos de FPA, el ganador es el que hace la oferta más alta y paga su propia oferta como precio, dado que la misma sea mayor al precio de reserva<sup>7</sup>. Repasaremos nuevamente cada modelo teórico dentro del paradigma

---

<sup>7</sup>En una licitación el ganador es el oferente que realizó la menor oferta y obtiene dicho valor por el bien o servicio adjudicado. El análisis es análogo al de subastas con los cambios apropiados en los signos.

de valuaciones privadas con un poco más de detalle para poder explicar luego de manera más transparente cómo formular el modelo econométrico correspondiente.

#### 4.1. El modelo IPV simétrico

En este paradigma los postores obtienen sus valuaciones privadas en forma independiente de una distribución común,  $F(., \dots, .)$ . En particular, tenemos que  $F(., \dots, .) = F(\cdot)^n$ . El BNE es simétrico, dado que el juego es simétrico. Por lo tanto cada jugador utiliza la misma estrategia,  $s(\cdot) : [\underline{v}, \bar{v}] \rightarrow [\underline{b}, \bar{b}]$ . Cuando  $n \geq 2$  hay solución única para el BNE, en este caso, de las condiciones de primer orden (CPO) se tiene que,

$$1 = [v_i - s(v_i)](n-1) \frac{f(v_i)}{F(v_i)} \frac{1}{s'(v_i)},$$

sujeito a la condición de frontera  $s(\underline{v}) = \underline{v}$ , es decir el jugador con la menor valuación hace la oferta más baja. Luego la solución puede escribirse como,

$$b_i = s(v_i, F, n) \equiv v_i - \frac{1}{[F(v_i)]^{n-1}} \int_{\underline{v}}^{v_i} [F(u)]^{n-1} du. \quad (3)$$

Este modelo ha sido ampliamente estudiado (ver Riley y Samuelson, 1981). Se han obtenido varios resultados teóricos importantes a partir de este modelo (*Revenue Equivalence Theorem Vickrey, 1961, Optimal Reserve Price*)<sup>8</sup>.

---

<sup>8</sup>El *Revenue Equivalence Theorem* establece que en el paradigma IPV las subastas ascendentes, las descendentes y las de sobre cerrado de primer y segundo precio tienen asociadas la misma recaudación esperada para el vendedor.

## 4.2. El modelo APV simétrico

Una debilidad del modelo IPV anterior es el supuesto restrictivo de independencia de las valuaciones. Un marco más general es por lo tanto el que nos brinda el modelo APV. En este caso los jugadores obtienen sus valoraciones privadas de una distribución conjunta y afiliada  $F(\cdot, \dots, \cdot)$ . Sea  $y_i = \max_{j \neq i} v_j$ , la estrategia de equilibrio satisface

$$s'(v_i) = [v_i - s(v_i)] \frac{f_{y_1|v_1}(v_i|v_i)}{F_{y_1|v_1}(v_i|v_i)},$$

para todo  $v_i \in [\underline{v}, \bar{v}]$ , sujeto a la condición de frontera  $s(\underline{v}) = \underline{v}$ .  $F_{y_1|v_1}(\cdot|\cdot)$  denota la distribución condicional de  $y_1$  dado  $v_1$  y  $f_{y_1|v_1}(\cdot|\cdot)$  es la densidad condicional respectiva. El índice "1" hace referencia a cualquier jugador dado que todos son ex-ante idénticos. De Milgrom y Weber (1982) sabemos que la solución es

$$b_i = s(v_i, F) = v_i - \int_{\underline{v}}^{v_i} L(\alpha|v_i) d\alpha, \quad (4)$$

con  $L(\alpha|v_i) = \exp\left[-\int_{\alpha}^{v_i} \frac{f_{y_1|v_1}(u|u)}{F_{y_1|v_1}(u|u)} du\right]$ .

Este modelo, con supuestos más realistas que el modelo IPV tiene también un amplio potencial para aplicaciones empíricas.

## 4.3. El modelo IPV asimétrico

Una de las limitaciones de los modelos simétricos es el supuesto que todos los jugadores son ex-ante idénticos, lo cual no se condice con muchas situaciones que uno ve en los datos empíricos. Para simplificar la exposición suponemos que hay dos tipos de jugadores, el modelo puede generalizarse fácilmente a una

situación con más de dos tipos (para un caso con tres tipos de jugadores ver Aryal y Gabrielli, 2013). Por ejemplo el grupo 1 puede ser el que contiene a los jugadores mejor informados o los de mayor tamaño, etc. El grupo 0 junta a los otros participantes. Sea  $n_j$  el número de jugadores de tipo  $j$ ,  $j = 0, 1$  con  $n_0 + n_1 = n \geq 2$ . El modelo asume que dentro de cada grupo o tipo los jugadores son ex-ante idénticos, es decir que sus valuaciones son variables aleatorias iid provenientes de  $F_j(\cdot)$ . Por lo tanto,  $F(\cdot, \dots, \cdot) = F_1(\cdot)^{n_1} F_0(\cdot)^{n_0}$ . Cada una de estas distribuciones son de conocimiento común y tienen un soporte común,  $[\underline{v}, \bar{v}]$ .

Las estrategias de equilibrio  $s_0(\cdot)$  y  $s_1(\cdot)$  por lo general no tienen solución analítica. No obstante estas estrategias satisfacen el siguiente sistema de ecuaciones diferenciales,

$$\begin{aligned} s_1'(v_{1i}) &= [v_{1i} - b_{1i}] \left[ (n_1 - 1) \frac{f_1(v_{1i})}{F_1(v_{1i})} + n_0 \frac{f_0(s_0^{-1}(b_{1i}))}{F_0(s_0^{-1}(b_{1i}))} \frac{s_1'(v_{1i})}{s_1'(s_0^{-1}(b_{1i}))} \right] \\ s_0'(v_{0i}) &= [v_{0i} - b_{0i}] \left[ (n_0 - 1) \frac{f_0(v_{0i})}{F_0(v_{0i})} + n_1 \frac{f_1(s_1^{-1}(b_{0i}))}{F_1(s_1^{-1}(b_{0i}))} \frac{s_0'(v_{0i})}{s_1'(s_1^{-1}(b_{0i}))} \right], \end{aligned}$$

sujeto a las condiciones de frontera  $s_0(\underline{v}) = s_1(\underline{v}) = \underline{v}$  y  $s_1(\bar{v}) = s_0(\bar{v})$ .

Dado que el sistema no tiene solución analítica, hay que recurrir a métodos numéricos lo que representa la mayor dificultad asociada a estos modelos.

#### 4.4. El modelo APV asimétrico

En este caso y bajo el supuesto de que hay dos tipos de jugadores este modelo considera que el vector  $(v_{11}, \dots, v_{1n_1}, v_{01}, \dots, v_{0n_0})$  se distribuye de acuerdo a la función conjunta  $F(\cdot)$ . Esta función se asume intercambiable en sus

primeros  $n_1$  componentes como también en sus últimos  $n_0$  componentes. La intuición detrás de este esquema probabilístico es que dentro de un mismo tipo o grupo los jugadores son idénticos ex-ante. Más aún, como  $F(\cdot)$  es afiliada hay dependencia general entre las valuaciones privadas. El trabajo de Campo et al. (1998) caracteriza el sistema de ecuaciones diferenciales que define el equilibrio de este juego. Al igual que en el caso anterior este sistema carece de solución analítica.

#### 4.5. Métodos de estimación

Dado que una subasta contiene solamente un puñado de observaciones, en la estimación de estos modelos típicamente se utiliza más de una subasta. Definamos  $\ell$  como el índice de la subasta,  $\ell = 1, \dots, L$ . Por lo tanto la estructura del modelo es  $[U_\ell(\cdot), F_\ell(\cdot)]$ . Pueden incorporarse características observables de las subastas,  $X_\ell$ , al modelo, en este caso se tiene que  $U_\ell(\cdot) = U(\cdot, X_\ell)$  y  $F_\ell(\cdot) = F(\cdot | X_\ell)$ .

Una primera clasificación que podemos hacer para la estimación estructural de modelos de subastas distingue entre métodos directos y métodos indirectos. A su vez estos métodos pueden ser paramétricos si los elementos estructurales se suponen conocidos excepto por un número finito de parámetros. También se pueden utilizar técnicas no paramétricas para la estimación en donde no se hacen supuestos sobre la forma funcional de los elementos estructurales.

#### 4.5.1. Métodos directos

Estos métodos fueron los primeros que se desarrollaron en la literatura y están basados principalmente en modelos econométricos paramétricos. El punto de partida es la especificación de la función de distribución de las valoraciones privadas,  $F(\cdot)$ , de manera de poder estimar los parámetros que caracterizan dicha distribución. En este grupo de métodos hay dos grandes clases de procedimientos. En primer lugar se encuentra la metodología desarrollada por Paarsch (1992) y Donald y Paarsch (1993) que consiste en un método enteramente paramétrico que utiliza Máxima Verosimilitud (MV) para el cómputo de las estrategias de equilibrio. Donald y Paarsch (1993) reconocen que esta metodología es altamente demandante en términos computacionales y por lo tanto solamente distribuciones sencillas pueden utilizarse en la práctica. En particular, el problema es que el extremo superior de la función de distribución de las ofertas depende del o de los parámetros a estimar (digamos  $\theta$ ) lo que deriva en una distribución no estándar para el estimador de MV<sup>9</sup>. Es básicamente este motivo lo que llevó a Donald y Paarsch (1993) a desarrollar el ya bien conocido pseudo estimador MV por partes (*piecewise pseudo ML estimator*). Este estimador requiere que la estrategia de equilibrio sea calculada de manera explícita, lo que usualmente se obtiene a partir de distribuciones paramétricas. Donald y Paarsch (1993) han establecido las propiedades asintóticas de este estimador, es decir han demostrado su consistencia y que la distribución del estimador converge a la tasa (paramétrica)  $\sqrt{L}$  a una distribución normal. Dada la dificultad asociada al cómputo de la estrategia de equilibrio, Marshall et al. (1994) proponen un conjunto de algoritmos numéricos para resolver de manera numérica la estrategia de

---

<sup>9</sup>La razón de este resultado es que al depender el soporte del parámetro de interés se está violando una de las condiciones de regularidad que necesitan los métodos de MV.

equilibrio de una subasta de primer precio asimétrica permitiendo el uso de distribuciones arbitrarias para las valoraciones privadas.

En un segundo trabajo, Donald y Paarsch (1996) estudian el problema de identificación del parámetro  $\theta$ , así como las propiedades del estimador de MV en el caso que el extremo superior del soporte dependa del parámetro desconocido. La distribución asintótica que obtienen para este estimador es no estándar lo que constituye una desventaja en la práctica. Ambos métodos basados en MV están sujetos a dos grandes limitaciones. En primer lugar, solo permiten la inclusión de heterogeneidad observable (de las subastas) a través de variables discretas. En segundo lugar, tenemos la gran complejidad para computar estos estimadores ya que las rutinas asociadas son demasiado complejas y requieren el cálculo de la estrategia de equilibrio y de su inversa para determinar el extremo superior de la distribución y la densidad de las ofertas.

Laffont et al. (1995) introducen una metodología computacionalmente más conveniente. El método se basa en el *Revenue Equivalence Theorem*. Los autores proponen un método basado en simulaciones que evita el cálculo de la estrategia de equilibrio y por lo tanto permite considerar especificaciones paramétricas más generales para las distribuciones de las valuaciones privadas. A este método se lo conoce como Mínimos Cuadrados No Lineales Simulados (SNLLS, por sus siglas en inglés) y se relaciona con los métodos propuestos por McFadden (1989) y por Pakes y Pollard (1989). El estimador SNLLS de Laffont et al. (1995) utiliza solamente a la oferta ganadora dado que los autores analizan una subasta oral descendente de berenjenas en su aplicación empírica. Los autores establecen las propiedades asintóticas del estimador, léase su consistencia y su normalidad a la tasa de convergencia  $\sqrt{L}$  para un número dado de simulaciones y mientras el número de subastas tiende a infinito. La principal ventaja de este método radica en que no es necesario el cómputo de la estrategia de equilibrio o de su inversa. Es por esto que este método es atractivo desde



un punto de vista computacional y puede ser utilizado con cualquier familia de distribuciones paramétricas para  $F(\cdot)$ <sup>10</sup>. Li y Vuong (1997) extendieron el método de SNLLS al caso en el cual se observan todas las ofertas, como sucede típicamente en las subastas a sobre cerrado.

#### **4.5.2. Métodos indirectos**

En contraste con los métodos directos, los procedimientos indirectos de estimación no utilizan la distribución de valuaciones privadas como punto de partida. El uso de estos métodos fue iniciado por Guerre et al. (2000) en su trabajo seminal sobre estimación estructural de subastas de primer precio. Los autores desarrollan un procedimiento enteramente no paramétrico. Esta alternativa está basada en una simple pero crucial observación, cada valoración privada puede expresarse como una función de la oferta correspondiente, de la distribución de las ofertas observadas y de su densidad. Esta función surge de las CPO del problema de optimización del jugador. Es por esto que, en contraposición a los métodos directos, el punto de partida de los procedimientos indirectos es la distribución de las ofertas (observadas) que se utiliza para recuperar la distribución de las valuaciones privadas (no observadas) sin tener que recurrir al cómputo de la estrategia que conforma el BNE o su función inversa de manera explícita.

La principal ventaja de este tipo de métodos es por lo tanto que no se requiere resolver la ecuación diferencial que caracteriza al equilibrio bayesiano de Nash, y, en consecuencia no es necesaria una solución analítica del sistema de ecuaciones. Más aún, dada la naturaleza no paramétrica del método bajo consideración, el procedimiento no está sujeto a supuestos paramétricos y por

---

<sup>10</sup>En la aplicación empírica los autores utilizan una distribución log-normal.

ende no restringe a priori a la función  $F(\cdot)$  a pertenecer a alguna clase de distribuciones específica. Otra ventaja asociada a este esquema es que permite la derivación de resultados importantes sobre identificación y de restricciones testeables provenientes de la teoría subyacente que pueden utilizarse para validar el modelo<sup>11</sup>.

Utilizando como base la ecuación que define la función inversa de la estrategia de equilibrio, los autores muestran que el modelo está identificado de manera no paramétrica. Hay una serie de trabajos en la literatura (de estos autores y de otros) que consideran de manera similar otros modelos de subastas, como por ejemplo modelos con valores privados afiliados, modelos con jugadores asimétricos, modelos dinámicos y modelos con postores que son adversos al riesgo.

El método de Guerre et al. (2000) naturalmente se implementa a través de un procedimiento en dos etapas. En un primer paso, una muestra de pseudo valuaciones privadas es obtenida a través del uso de estimadores no paramétricos para la distribución y densidad de las ofertas observadas. Esta muestra se utiliza luego en una segunda etapa para estimar también de manera no paramétrica la densidad de las valuaciones privadas de los jugadores. Los autores muestran que el estimador propuesto tiene características estadísticas deseables tales como consistencia uniforme y que logran una tasa de convergencia óptima al elegir valores adecuados de los anchos de banda (*bandwidths*).

---

<sup>11</sup>Estudiaremos el método propuesto por Guerre et al. (2000) en detalle más adelante.

## 5. Estimación estructural del modelo IPV simétrico

### 5.1. Modelo teórico e identificación

El trabajo de Guerre et al. (2000) (GPV, en adelante) introdujo el uso de métodos indirectos para estimar modelos de subastas de primer precio dentro del paradigma de valuaciones privadas independientes.

En particular, los autores proponen un procedimiento indirecto en dos etapas para estimar la función de distribución de las valuaciones privadas de los jugadores a partir de las ofertas observadas. El método propuesto no requiere supuestos paramétricos ni tampoco el cálculo de la estrategia de equilibrio bayesiana de Nash. La idea fundamental detrás del procedimiento es que cada valuación privada puede expresarse como una función de la oferta correspondiente y de la distribución acumulada y densidad de ofertas observadas. En el primer paso se construye una muestra de pseudo valuaciones privadas mediante el uso de estimadores *kernels* para la distribución y densidad de las ofertas. En el segundo paso, la pseudo muestra se utiliza para estimar no paramétricamente la función de densidad de las valuaciones privadas de los individuos. En particular, el estimador propuesto es uniformemente consistente y óptimo ya que los autores demuestran que alcanza la mejor tasa de convergencia posible cuando se estima la densidad de los valores privados (no observados) a partir de ofertas (observadas).

Los resultados de este trabajo son importantes por varios motivos. Primero, las conclusiones de política basadas en procedimientos enteramente no paramétricos son robustas a posibles errores de especificación de la distribución subyacente. Más aún, dado que la teoría económica impone algunas restricciones sobre la distribución de las ofertas observadas, se puede testear en principio la validez del modelo teórico utilizado para caracterizar

el problema sin tener que recurrir a (fuertes) supuestos paramétricos sobre los elementos estructurales. En segundo lugar y desde un punto de vista estadístico, el estimador propuesto utiliza técnicas no paramétricas en cada etapa o paso del método estructural de estimación. La contribución estadística principal del trabajo es la derivación de la mejor tasa de convergencia (uniforme) para el estimador no paramétrico de la función de densidad de las valuaciones (no observables) privadas a partir de ofertas (observables). Para esto se utiliza la teoría minimax desarrollada por Ibragimov y Has'minskii (1981). Adicionalmente los autores muestran que mediante la elección de anchos de banda (*bandwidths*) apropiados, el estimador en dos etapas alcanza la tasa óptima de convergencia. En tercer lugar, y teniendo en cuenta aspectos computacionales, el estimador propuesto evita una serie de dificultades típicamente asociadas a los métodos directos utilizados previamente en el análisis empírico de subastas. De hecho, dado que el método es indirecto y enteramente no paramétrico, el procedimiento no requiere calcular en forma numérica la oferta de equilibrio ni tampoco requiere un algoritmo iterativo de optimización. Estos aspectos simplifican considerablemente la implementación de este tipo de procedimientos.

El modelo de referencia que utiliza GPV es el de valores privados independientes (IPV) explicado en la sección . Como es bien sabido, este modelo tiene solución analítica (Riley y Samuelson, 1981). Específicamente, para el caso de dos o más jugadores,  $N \geq 2$ , y con  $p_0$  que representa el precio de reserva,

$$b_i = s(v_i, F, N, p_0) \equiv v_i - \frac{1}{(F(v_i))^{N-1}} \int_{p_0}^{v_i} (F(u))^{N-1} du \quad (5)$$

si  $v_i \geq p_0$ . Si  $v_i < p_0$ , entonces  $b_i$  puede tomar cualquier valor menor que el precio de reserva  $p_0$ .

Esta estrategia se obtiene resolviendo la siguiente ecuación diferencial de primer orden en  $s(\cdot)$

$$1 = \left( v_i - s(v_i) \right) (N - 1) \frac{f(v_i)}{F(v_i)} \frac{1}{s'(v_i)}, \quad (6)$$

con condición de frontera  $s(p_0) = p_0$ . La estrategia de equilibrio (5) es estrictamente creciente en  $v_i$  en el intervalo  $[p_0, \bar{v}]$  y expresa a la oferta de equilibrio como una función de las valuaciones privadas, de la distribución de dichas valuaciones, del número de jugadores y del precio de reserva.

Por lo general, las ofertas se observan en los datos mientras que las valuaciones privadas no. El marco teórico precedente está estrechamente vinculado con un modelo econométrico estructural. En particular, dado que  $b_i$  es una función de  $v_i$ , la cual es aleatoria y se distribuye de acuerdo a la función  $F(\cdot)$ , se sigue que  $b_i$  es también aleatoria y se distribuye de acuerdo a la función  $G(\cdot)$  (digamos). Esta simple observación es la base del análisis estructural de datos de subastas.

GPV proponen un método de estimación indirecto ya que comienzan estimando la distribución de las ofertas observadas para luego construir un estimador de la distribución de los valores privados. Para el resto del análisis suponemos que el precio de reserva es no vinculante (*nonbinding*), i.e.  $p_0 = \underline{v}$  de manera que  $s(\underline{v}) = \underline{v}$ . Este supuesto puede ser fácilmente relajado, pero a los efectos de una exposición más transparente lo mantendremos y luego indicaremos los cambios necesarios que deben hacerse si se abandona.

Un aspecto fundamental que subyace en la estimación estructural es poder establecer si los elementos estructurales pueden ser identificados a partir de las observaciones disponibles. Dado que estamos suponiendo un precio de reserva no vinculante, el número de postores potenciales  $N$  es igual a la cantidad de participantes actuales, es por ello que tanto  $N$  como  $b_i$ ,  $i = 1, \dots, N$  se conocen. Por lo tanto el único elemento estructural desconocido es la distribución latente de las valuaciones privadas,  $F(\cdot)$ , y de esta forma el problema de la identificación se reduce a determinar esta función de forma unívoca a partir de los datos.

El Teorema 1 de GPV resuelve el problema mencionado al establecer que la distribución  $F(\cdot)$ , de existir, es única. Adicionalmente, este teorema brinda condiciones necesarias y suficientes sobre la distribución  $G(\cdot)$  que denota la distribución conjunta de  $(b_1, \dots, b_N)$ , para la existencia de  $F(\cdot)$ . Como es usual, se restringe la atención a estrategias de equilibrio Bayesianas de Nash estrictamente crecientes y diferenciables<sup>12</sup>.

**Teorema 1 - GPV 2000:** *Sea  $N \geq 2$ . Sea  $\mathbf{G}(\cdot)$  una función que pertenece a la familia  $\mathcal{P}^N$  (conjunto de distribuciones de probabilidad continuas) con soporte  $[\underline{b}, \bar{b}]^N$ . Existe una distribución para las valuaciones privadas de los postores  $F(\cdot) \in \mathcal{P}$  de manera que  $\mathbf{G}(\cdot)$  es la distribución de las ofertas de equilibrio de una subasta a sobre cerrado de primer precio con valuaciones privadas independientes y un precio de reserva no vinculantes si y solo si*

C1:  $\mathbf{G}(b_1, \dots, b_N) = \prod_{i=1}^N G(b_i)$ .

C2: *La función  $\xi(\cdot, G, N)$  es estrictamente creciente en  $[\underline{b}, \bar{b}]$  y su inversa es diferenciable en  $[\underline{v}, \bar{v}] \equiv [\xi(\underline{b}, G, N), \xi(\bar{b}, G, N)]$ .*

*Más aún, si  $F(\cdot)$  existe, es única con soporte  $[\underline{v}, \bar{v}]$  y satisface  $F(v) = G(\xi^{-1}(v, G, N))$  para todo  $v \in [\underline{v}, \bar{v}]$ . Adicionalmente,  $\xi(\cdot, G, N)$  es la cuasi inversa de la estrategia de equilibrio en el siguiente sentido  $\xi(b, G, N) = s^{-1}(b, F, N)$  para todo  $b \in [\underline{b}, \bar{b}]$ .*

Este resultado está basado en el hecho de que la primera derivada  $s'(\cdot)$  y la distribución  $F(\cdot)$  y su densidad  $f(\cdot)$  pueden ser eliminadas simultáneamente de la ecuación diferencial (6) que define el equilibrio, a través de la introducción de la distribución de las ofertas  $G(\cdot)$  y de su densidad  $g(\cdot)$ . Específicamente, para cada  $b \in [\underline{b}, \bar{b}] = [\underline{v}, s(\bar{v})]$ , tenemos  $G(b) = \Pr(\tilde{b} \leq b) = \Pr(\tilde{v} \leq s^{-1}(b)) = F(s^{-1}(b)) = F(v)$ , donde en la última igualdad se ha utilizado,  $b = s(v)$ . Por lo tanto, se tiene que  $G(\cdot)$  es absolutamente continua con soporte  $[\underline{v}, s(\bar{v})]$  y densidad  $g(b) = f(v)/s'(v)$ , donde  $v = s^{-1}(b)$ . Tomando el ratio obtenemos,

---

<sup>12</sup>Tal como lo advierten los autores, se asume que las condiciones de segundo orden se cumplen.

Por lo tanto el BNE está completamente caracterizado por la CPO (6).

$g(b)/G(b) = (1/s'(v))f(v)/F(v)$ . Por lo tanto, la ecuación diferencial (6) puede escribirse como,

$$v_i = \xi(b_i, G, I) \equiv b_i + \frac{1}{N-1} \frac{G(b_i)}{g(b_i)}. \quad (7)$$

La importancia de este teorema radica en varios aspectos. En primer lugar, muestra que el modelo teórico impone algunas restricciones sobre la distribución de las ofertas. La condición C1 establece que las ofertas son independientes e idénticamente distribuidas (iid) de acuerdo a la función  $G(\cdot)$ . La condición C2 dice que, dado  $N$ , la distribución  $\mathbf{G}(\cdot)$  de la ofertas observadas puede racionalizarse a través de una distribución de valuaciones privadas  $F(\cdot)$  solo si  $\xi(\cdot, G, N)$  es estrictamente creciente. En segundo lugar, asumiendo que los postores se comportan de acuerdo al modelo teórico expuesto, el Teorema 1 establece que la distribución de las valoraciones privadas de los jugadores,  $F(\cdot)$ , está identificada a partir de la distribución de las ofertas. En particular este resultado revela que no se requieren especificaciones paramétricas a priori para lograr la identificación del modelo estructural.

Finalmente, es de utilidad remarcar que la función  $\xi(\cdot, G, N)$  está completamente determinada a partir del conocimiento de  $G(\cdot)$  y  $N$ . Dado que  $\xi(\cdot, G, N)$  es la cuasi inversa de  $s(\cdot, F, N)$ , no se tiene que resolver en forma explícita la ecuación diferencial (6) ni tampoco se necesita aplicar integración numérica en (5) para determinar  $s(\cdot, F, N)$ , la estrategia de equilibrio de los participantes.

Para analizar modelos más generales introducimos heterogeneidad observable en las subastas. El vector  $X_\ell$  denota las características observables relevantes del  $\ell$ -ésimo objeto subastado y la variable  $N_\ell$  denota el número de jugadores en dicha subasta<sup>13</sup>. La distribución condicional de las valuaciones privadas en la  $\ell$ -ésima subasta dado  $X_\ell$  es  $F(\cdot | X_\ell, N_\ell)$ . De manera similar, la

<sup>13</sup>Dado que  $F(\cdot)$  es de conocimiento común en el modelo teórico, el vector  $X_\ell$  también se asume de conocimiento común para todos los participantes (vendedor y compradores). La justificación de esto radica en el hecho de que las características de los objetos subastados se

distribución condicional de las ofertas observadas dado  $X_\ell$  es  $G(\cdot | X_\ell, N_\ell)$ . Por lo tanto las ecuaciones que definen el equilibrio pueden expresarse como

$$\begin{aligned} B_{p\ell} &\equiv s(V_{p\ell}, X_\ell, N_\ell) = V_{p\ell} - \frac{1}{F(V_{p\ell} | X_\ell, N_\ell)^{N_\ell-1}} \int_{\underline{v}_\ell}^{V_{p\ell}} F(v | X_\ell, N_\ell)^{N_\ell-1} dv, \\ V_{p\ell} &\equiv \xi(B_{p\ell}, X_\ell, N_\ell) = B_{p\ell} + \frac{1}{N_\ell - 1} \frac{G(B_{p\ell} | X_\ell, N_\ell)}{g(B_{p\ell} | X_\ell, N_\ell)}, \end{aligned} \quad (8)$$

donde  $\underline{v}_\ell \equiv \underline{v}(X_\ell, N_\ell)$  es el extremo inferior del soporte de  $F(\cdot | X_\ell, N_\ell)$  y  $g(\cdot | \cdot, \cdot)$  es la densidad de  $G(\cdot | \cdot, \cdot)$ .

## 5.2. Método de estimación

La idea básica detrás del procedimiento de estimación es sencilla. Si uno conociera la distribución  $G(\cdot)$  y la densidad  $g(\cdot)$ , entonces podría utilizar (7) para recuperar  $v_i$  con el objetivo de estimar luego  $f(\cdot)$ . No obstante, en la práctica  $G(\cdot)$  y  $g(\cdot)$  son desconocidas, pero pueden ser estimadas a partir de las ofertas observadas. Esto sugiere implementar un mecanismo en dos etapas. En la primera etapa se construye una muestra de pseudo valuaciones privadas a partir de (7) o (8) mediante el uso de estimadores no paramétricos para la distribución y densidad de las ofertas. En el segundo paso se utiliza la muestra de pseudo valuaciones privadas para estimar no-paraméricamente la densidad de los valores privados de los oferentes. A continuación mostramos como adaptar esta metodología a subastas heterogéneas. También veremos los

---

encuentran descriptos en folletos gratuitos, por ejemplo. Se asume que ninguna característica relevante está omitida en  $X_\ell$ . Por lo tanto la heterogeneidad no observable solamente proviene de los diferentes valores privados de los individuos, que son variables aleatorias no observadas en términos del modelo estructural.



supuestos sobre los *kernels* y los anchos de banda (*bandwidths*) que se necesitan para definir el estimador y obtener optimalidad.

Notar que (8) puede reescribirse como

$$V_{p\ell} \equiv \xi(B_{p\ell}, X_\ell, N_\ell) = B_{p\ell} + \frac{1}{N_\ell - 1} \frac{G(B_{p\ell}, X_\ell, N_\ell)}{g(B_{p\ell}, X_\ell, N_\ell)}, \quad (9)$$

con  $G(b, x, n) = G(b|x, n) f_m(x, n) = \int_{\underline{b}(x)}^b g(u, x, n) du$ .

Por lo tanto, utilizando las observaciones  $\{(B_{p\ell}, X_\ell, N_\ell); p = 1, \dots, N_\ell, \ell = 1, \dots, L\}$ , en la primera etapa se estima el siguiente ratio  $\psi(\cdot, \cdot, \cdot) = G(\cdot, \cdot, \cdot)/g(\cdot, \cdot, \cdot)$  a partir de  $\tilde{\psi} = \tilde{G}/\tilde{g}$ , donde

$$\tilde{G}(b, x, n) = \frac{1}{L h_G^d} \sum_{\ell=1}^L \frac{1}{N_\ell} \sum_{p=1}^{N_\ell} \mathbf{I}(B_{p\ell} \leq b) K_G \left( \frac{x - X_\ell}{h_G}, \frac{n - N_\ell}{h_{GN}} \right), \quad (10)$$

$$\tilde{g}(b, x, n) = \frac{1}{L h_g^{d+1}} \sum_{\ell=1}^L \frac{1}{N_\ell} \sum_{p=1}^{N_\ell} K_g \left( \frac{b - B_{p\ell}}{h_g}, \frac{x - X_\ell}{h_g}, \frac{n - N_\ell}{h_{gN}} \right), \quad (11)$$

Los parámetros  $h_G$ ,  $h_{GN}$ ,  $h_g$ , y  $h_{gN}$  son *bandwidths*, y  $K_G$  y  $K_g$  son *kernels* con soporte acotado. La ecuación (9) revela que naturalmente las valuaciones privadas  $V_{p\ell}$  pueden estimarse como,

$$\tilde{V}_{p\ell} = B_{p\ell} + \frac{1}{N_\ell - 1} \tilde{\psi}(B_{p\ell}, X_\ell, N_\ell). \quad (12)$$

Luego estas estimaciones se pueden usar en una segunda etapa para recuperar la distribución condicional  $f(v|x)$ .

El problema que existe es que los estimadores *kernels* no paramétricos propuestos están asintóticamente sesgados en los extremos del soporte de  $(B, X, N)$ , por lo tanto no es conveniente utilizar  $\tilde{\psi}$ . Esta observación lleva a los autores a proponer un mecanismo de descarte (*trimming*) para poder evitar este sesgo que no desaparece con el tamaño de la muestra.

Por lo tanto en lugar de (12) se utiliza *trimming* en la zona cercana a los extremos. No obstante, primero hay que estimar los extremos del soporte de

$(B, X, N)$ , que son desconocidos. Formalmente, Guerre et al. (2000) proponen estimar los extremos del soporte de la densidad condicional de  $B$  dado  $(X, N) = (x, n)$  utilizando,

$$\begin{aligned}\hat{b}(x, n) &= \sup\{B_{p\ell}, p = 1, \dots, N_\ell, \ell = 1, \dots, L; X_\ell \in \pi_{k_1, \dots, k_d}, N_\ell = n\} \\ \underline{\hat{b}}(x, n) &= \underline{\hat{b}}(x) = \inf\{B_{p\ell}, p = 1, \dots, N_\ell, \ell = 1, \dots, L; X_\ell \in \pi_{k_1, \dots, k_d}\}. \quad (13)\end{aligned}$$

Dado que  $\underline{b}(x, n) = \underline{v}(x)$  es independiente de  $i$ , no es necesario restringir el ínfimo en (13) a ofertas tal que  $N_\ell = n$ . Denotemos por  $\hat{S}_n(G)$  al soporte de la función  $G$  estimado a partir de una muestra de tamaño  $n$ . El estimador para  $S_n(G)$  es  $\hat{S}_n(G) \equiv \{(b, x) : b \in [\underline{\hat{b}}(x, n), \hat{b}(x, n)], x \in [x, \bar{x}]\}$ .

Para entender como los autores manejan el descarte de observaciones en la cercanía de los extremos (*trimming*) hay que definir cuáles son las observaciones “cercanas a los extremos”. Básicamente, para cada  $n$  se descartan observaciones  $(B_{p\ell}, X_\ell, N_\ell)$  con  $N_\ell = n$  que estén ubicadas cerca de los extremos del soporte (una distancia de  $2h$ ). Esta cercanía se define por medio de una distancia que es función de los parámetros de suavización. En lugar de la ecuación (12) se definen los valores privados de la siguiente manera,

$$\hat{V}_{p\ell} = B_{p\ell} + \frac{1}{N_\ell - 1} \hat{\Psi}(B_{p\ell}, X_\ell, N_\ell), \quad (14)$$

donde

$$\hat{\Psi}(b, x, n) \equiv \begin{cases} \tilde{\Psi}(b, x, n) & \text{si } (b, x) + S(2h_G) \subset \hat{S}_n(G) \text{ y} \\ & (b, x) + S(2h_g) \subset \hat{S}_n(G), \\ +\infty & \text{en otro caso.} \end{cases} \quad (15)$$

En la segunda etapa se utiliza entonces la *pseudo muestra*  $\{(\hat{V}_{p\ell}, X_\ell), p = 1, \dots, N_\ell, \ell = 1, \dots, L\}$ , para estimar no-paramétricamente la densidad  $f(v|x)$  a través de  $\hat{f}(v|x) = \hat{f}(v, x) / \hat{f}(x)$ , donde

$$\hat{f}(v, x) = \frac{1}{Lh_f^{d+1}} \sum_{\ell=1}^L \frac{1}{I_\ell} \sum_{p=1}^{I_\ell} K_f \left( \frac{v - \hat{V}_{p\ell}}{h_f}, \frac{x - X_\ell}{h_f} \right), \quad (16)$$

$$\hat{f}(x) = \frac{1}{Lh_X^d} \sum_{\ell=1}^L K_X \left( \frac{x - X_\ell}{h_X} \right), \quad (17)$$

$h_f$  y  $h_X$  son *bandwidths*, y  $K_f$  y  $K_X$  son *kernels*.

Debido a que los *kernels* tienen soporte compacto, la contribución al estimador,  $\hat{f}(v, x)$ , de valuaciones privadas con valor infinito es nula. Por lo tanto, la ecuación (15) se interpreta como la regla de descarte o el *trimming* para observaciones  $(B_{p\ell}, X_\ell, N_\ell)$  que están demasiado cerca de los extremos del soporte estimado  $\hat{S}_{N_\ell}(G)$ .

Los siguientes supuestos establecen las opciones de *bandwidths* y *kernels*. El supuesto A3 es estándar en la literatura. El orden de los *kernels* se elige de acuerdo al grado de diferenciabilidad de las funciones estimadas. En particular  $K_g$  se elige de orden  $R + 1$  dado que la densidad  $g(\cdot, \cdot, \cdot)$  admite hasta  $R + 1$  derivadas parciales continuas.

**Supuesto A3:**

- (i) Los kernels  $K_G(\cdot, \cdot), K_g(\cdot, \cdot, \cdot), K_f(\cdot, \cdot), K_X(\cdot)$  son simétricos con soporte acotado y con derivadas de segundo orden con respecto a los argumentos continuos.
- (ii)  $\int K_G(x, 0)dx = 1, \int K_g(b, x, 0)dbdx = 1, \int K_f(v, x)dvdv = 1, \int K_X(x)dx = 1.$
- (iii)  $K_G(\cdot, 0), K_g(\cdot, \cdot, 0), K_f(\cdot, \cdot), K_X(\cdot)$  son de orden  $R + 1, R + 1, R,$  y  $R + 1,$  respectivamente. Por lo tanto, los momentos de orden estrictamente menor al orden establecido son iguales a cero.

**Supuesto A4:**

- (i) A medida que  $L \rightarrow \infty,$  los *bandwidths* “discretos”  $h_{G\ell}$  y  $h_{g\ell}$  tienden a cero.

(ii) Los bandwidths “*continuos*”  $h_G$ ,  $h_g$ ,  $h_f$ , y  $h_X$  son de la siguiente forma:

$$h_G = \lambda_G (\log L/L)^{1/(2R+d+2)}, \quad h_g = \lambda_g (\log L/L)^{1/(2R+d+3)}, \\ h_f = \lambda_f (\log L/L)^{1/(2R+d+3)}, \quad h_X = \lambda_X (\log L/L)^{1/(2R+d+2)},$$

donde las constantes  $\lambda$ 's son estrictamente positivas.

(iii) El bandwidth “*de la frontera*” tiene la siguiente forma  $h_\partial = \lambda_\partial (\log L/L)^{1/(d+1)}$  si  $d > 0$ .

Dados estos supuestos, los estimadores propuestos en (10), (11) y (17) para  $G(\cdot, \cdot, \cdot)$ ,  $g(\cdot, \cdot, \cdot)$  y  $f(\cdot)$ , respectivamente, convergen de manera uniforme a la mejor tasa posible. Si en cambio, las valuaciones privadas se observaran, el *bandwidth* óptimo para estimar  $f(\cdot, \cdot)$  sería de orden  $(\log L/L)^{1/(2R+d+1)}$ , que es asintóticamente menor que la tasa a la que converge  $h_f$  establecida en A4-(ii). Por lo tanto esta opción de *bandwidth* implica “*oversmoothing*” (suavizar de más), y en este sentido sería sub-óptimo. No obstante, las valuaciones privadas no se observan y por lo tanto dadas las tasas óptimas para  $h_G$ ,  $h_g$  y  $h_X$  (es decir los estimadores de la primera etapa), la elección que hacen Guerre et al. (2000) para  $h_f$  es la única posible para alcanzar la tasa de convergencia óptima establecida en el Teorema 2.

Podemos resumir diciendo lo siguiente. El trabajo de Guerre et al. (2000) muestra que la distribución subyacente de las valuaciones privadas independientes en un modelo de subastas de primer precio a sobre cerrado, ya sea con precio de reserva vinculante o no, está identificada a partir de las ofertas observadas y del número de jugadores sin tener que recurrir a supuestos paramétricos. Más aún, usando la teoría min-max, los autores establecen la mejor tasa de convergencia uniforme para estimar la distribución latente de las valuaciones privadas a partir de variables observables. También proponen un estimador no paramétrico en dos etapas que es computacionalmente conveniente.

En realidad, los resultados de este trabajo pueden extenderse más allá del mecanismo de subasta analizado y del paradigma supuesto. Por ejemplo,

se puede utilizar esta metodología para el paradigma de valuaciones privadas independientes pero cuyo mecanismo de asignación sea el de subastas Holandesas donde solamente se observa la oferta ganadora. Alternativamente el método pueden extenderse para el caso de valuaciones privadas interdependientes (afiliadas) que constituyen el modelo más general de subastas que puede ser identificado (Laffont y Vuong, 1996). En términos generales, dado que el método propuesto evita la determinación de la estrategia de equilibrio, este procedimiento es especialmente conveniente cuando dicha estrategia no tiene una forma analítica cerrada como ocurre cuando el juego es asimétrico.

Finalmente, un aspecto importante de los resultados de identificación obtenidos es que los mismos nos proveen condiciones necesarias y suficientes para la existencia de una distribución latente que puede “racionalizar” la distribución observada de las ofertas. Dado que estas restricciones sobre las variables observables vienen impuestas desde el modelo teórico, pueden constituir la base para realizar un test sobre dicha teoría. Hay dos clases de restricciones. La primera está relacionada con la independencia de las ofertas y se refiere al paradigma (IPV o CV) y podría corroborarse con algún test sobre independencia. El segundo tipo de restricciones tiene que ver con la monotonicidad de la función estimable lo que puede utilizarse para chequear si los jugadores adoptan la estrategia simétrica de equilibrio Bayesiano de Nash.

Aryal et al. (2017) extienden la metodología GPV a un procedimiento semi paramétrico en dos etapas. En particular estos autores proponen un método semi paramétrico en dos etapas basado en restricciones de momento para estimar un conjunto finito de parámetros que caracterizan a la densidad de las valuaciones privadas en subastas de primer precio. La utilización de una metodología semi paramétrica permite acomodar un mayor número de variables (características observables de las subastas) en el análisis, a la vez que se evita el problema usual de la maldición de la dimensionalidad y los

problemas de estimación en los extremos asociados a la estimación mediante *kernels* del tipo Parzen-Rosenblatt como el que utilizan GPV. En el primer paso de Aryal et al. (2017) se estima la distribución y densidad de las ofertas observadas utilizando regresiones por polinomios locales (metodología no paramétrica que generaliza a los *kernels*, ver Fan y Gijbels, 1996) que permite recuperar una muestra de (pseudo) valuaciones privadas. En el segundo paso se aplica el método generalizado de momentos (GMM) para estimar los parámetros de interés. Los autores muestran que este estimador es consistente, converge a las tasa paramétrica ( $\sqrt{n}$ ) y es asintóticamente normal. Si bien, en la literatura empírica de estimación de subastas varios autores han utilizado métodos semi paramétricos para estimar los modelos, el trabajo de Aryal et al. (2017) es el primero en establecer formalmente las propiedades asintóticas del estimador. El trabajo también contiene un experimento Monte Carlo en donde se muestra que este estimador también funciona muy bien en muestras finitas tanto en términos de estimar la densidad como en la elección del precio de reserva que maximiza la recaudación.

## **6. Detectando colusión dentro del Enfoque estructural**

Las subastas y las licitaciones son mecanismos comúnmente utilizados en la práctica tanto por el sector público como por el sector privado para asignar bienes y servicios. Desde un punto de vista teórico, típicamente se asume que existe competencia entre los participantes para obtener conclusiones sobre eficiencia. En otras palabras, el diseño óptimo del mecanismo asume que no hay colusión o comportamiento cooperativo entre los participantes. No obstante, tanto subastas como licitaciones son susceptibles a la presencia de cárteles, donde los participantes coluden en detrimento de la competencia.

La evidencia sugiere que este tipo de prácticas son más frecuentes de lo que a uno le gustaría. Ofertas acordadas se han encontrado en varios mercados, como por ejemplo en la obra pública, en el mercado que provee leche para escuelas, en el mercado de estampillas postales, etc.; ver Asker (2008), Bajari (2001), Comanor y Schankerman (1976), Feinstein et al. (1985), Harrington (2008), Lang y Rosenthal (1991), Pesendorfer (2000), Porter y Zona (1993), Porter y Zona (1999) y en mercados financieros como por ejemplo el de bonos municipales, entre otros. A pesar de que en los últimos años la imposición criminal de las leyes antimonopólicas haya desalentado la fijación de precios en algunos mercados no ha desalentado la colusión (Marshall y Meurer, 2001) y por lo tanto la importancia de cualquier método para detectar colusión no ha disminuido.

Por lo tanto, surgen las siguientes preguntas importantes: si hubiera colusión, ¿podemos detectarla a partir de los datos?, ¿podemos usar esta información para diseñar mejores reglas?. Dado que la colusión disminuye la recaudación del estado o aumenta el costo de las licitaciones que usualmente son financiadas a partir de impuestos distorsivos, es que estas prácticas derivan típicamente en mayor ineficiencia. Por lo tanto, la mayor recaudación que se destina a licitaciones debido a la presencia de colusión no es simplemente una transferencia de riqueza desde los contribuyentes a los cárteles. Luego, es importante y urgente poder detectar y detener estas prácticas.

Trabajos empíricos previos sobre colusión se han basado principalmente en información proveniente de juicios civiles para estimar el costo social de la colusión o han utilizado técnicas de estimación de forma reducida que ignoran las potenciales interacciones estratégicas entre los que coluden, lo cual lleva a errores de especificación del modelo y en consecuencia a resultados no confiables. No es una exageración decir que dichos datos son extremadamente difíciles de obtener, y aún cuando estén disponibles, en la mayoría de los casos es demasiado tarde. Por lo tanto es deseable tener métodos que permitan

determinar si los participantes de un mercado coluden basados en información pública sobre las ofertas correspondientes. Otra razón importante para estudiar la posible presencia de colusión en subastas o licitaciones radica en que en estos mecanismos los objetivos de maximizar recaudación (o minimizar costos) y eficiencia dependen crucialmente de la existencia de competencia entre los participantes (Ausubel y Milgrom, 2006).

La literatura sobre comportamiento competitivo en subastas y licitaciones ha desarrollado una serie de herramientas estadísticas para intentar captar si el comportamiento de los agentes económicos es consistente con las predicciones de dicha teoría. Al igual que en el caso de la estimación de modelos, las herramientas utilizadas para testear pueden clasificarse en dos grandes grupos. En primer lugar, existen los tests de forma reducida, como por ejemplo el test de Independencia Condicional y el Test de Simetría, desarrollados por Bajari y Ye (2003). Este tipo de procedimiento es particularmente útil cuando el analista no tiene información sobre quiénes son los jugadores que podrían estar actuando de manera no competitiva. En otras palabras, estos tests ayudan a identificar posibles candidatos cuyos comportamientos no se ajustan a lo que se esperaría en un ambiente competitivo.

En segundo lugar existen tests basados en el análisis estructural que intentan detectar comportamientos no consistentes con competencia a partir de la formulación de modelos teóricos que explican cómo las firmas forman sus ofertas (estratégicamente) en un ambiente competitivo y cómo lo harían en un ambiente no competitivo. En este sentido, Aryal y Gabrielli (2013) proponen un método simple que se implementa en dos etapas con el objetivo de testear si las ofertas observadas son consistentes con un modelo de colusión. En particular, los autores proponen testear la presencia de colusión en licitaciones públicas mediante un procedimiento no paramétrico con jugadores asimétricos en un modelo con costos privados independientes (paradigma IPV con jugadores



asimétricos)<sup>14</sup>. La idea del procedimiento es recuperar los costos privados subyacentes en un esquema competitivo y en un esquema colusivo. Puesto que, como se explica más adelante, para una oferta dada (y observada) el costo recuperado a partir de un modelo con colusión debe ser menor al recuperado en un ambiente competitivo (porque la colusión lo que hace es aumentar los *markups* o márgenes), la detección de la colusión puede postularse como testear por la presencia de dominancia estocástica de primer orden cuando se compara la distribución de los costos bajo un esquema competitivo con la que surgiría en un esquema colusivo.

Mostramos la intuición detrás de esta idea abstrayéndonos de la asimetría para facilitar la exposición, en la siguiente sección indicamos cómo adaptar los modelos cuando los jugadores son asimétricos. Sea  $F(\cdot)$  la distribución de los costos privados. Llamemos Modelo A al que utilizamos para describir un ambiente competitivo y Modelo B al correspondiente a un ambiente con colusión. Dado el argumento anterior sobre los márgenes de beneficio, si hay colusión entonces se tiene que  $Costo_A > Costo_B$  (en sentido estocástico) donde  $Costo_A$  es el costo recuperado bajo el supuesto de competencia y  $Costo_B$  es el obtenido a partir de un modelo con colusión. Luego, la hipótesis a testear puede escribirse como

$$F(t|competencia) \leq F(t|colusion) \quad \forall t.$$

Es decir, en presencia de colusión, la distribución (estimada) de los costos debe ser mayor (Dominancia Estocástica de Primer Orden) que si hay competencia. Naturalmente, surge el siguiente test:

$$\begin{array}{ll} H_0 : F^A = F^B & \text{Competencia} \\ H_1 : F^A \leq F^B & \text{Colusión} \end{array}$$

<sup>14</sup>Cuando lo que se analiza es un licitación y no una subasta, la variable no observada es típicamente el costo privado de cada jugador, que juega el mismo rol que la valoración privada en el contexto de subastas.

Dos tests usualmente utilizados para comparar distribuciones son el test de Rangos (Mann-Whitney-Wilcoxon) y el test de Kolmogorov-Smirnov. En este caso se utilizan estos mecanismos para testear las hipótesis y en definitiva discriminar entre Competencia y Colusión. En ambos casos rechazar la hipótesis nula implica evidencia en favor de colusión y no rechazarla implica evidencia en favor de un modelo competitivo.

Es importante modelar cómo se realizan las prácticas colusivas. Aryal y Gabrielli (2013) utilizan el esquema más sencillo posible, que se conoce en la literatura como colusión eficiente y básicamente implica que la firma con el menor costo es la que participa seriamente en la licitación y las otras firmas que pertenecen al cártel solo participan (si es que lo hacen) para proveer una oferta “fantasma” y de esa manera crear la apariencia de competencia. Pero, dado que la lógica de un test de hipótesis es similar a la lógica que se utiliza en el sistema judicial, es decir un acusado es inocente hasta tanto se encuentre evidencia significativa de que es culpable, la hipótesis nula del test postula competencia (i.e. ambas distribuciones son iguales), mientras que la hipótesis alternativa es lo que sucede cuando hay suficiente evidencia en contra de la igualdad de las distribuciones. En el presente caso, la hipótesis alternativa relevante es la que postula dominancia estocástica de primer orden (i.e. colusión). Como mencionamos más arriba, en el análisis estructural la relación entre el modelo teórico subyacente y el modelo econométrico resultante es un aspecto clave. Es por esto que a continuación detallamos los aspectos básicos de los modelos que se utilizan para describir un ambiente competitivo y uno que no lo es.

## 6.1. El modelo de competencia (Modelo A)

Un único bien indivisible es licitado (en cada momento del tiempo) a  $N \geq 2$  firmas o jugadores que se asumen neutrales al riesgo. Se consideran dos tipos de jugadores ( $k = 0, 1$ ) y en cada subconjunto existen  $n_k$  firmas de tipo  $k$ . Los jugadores tipo 1 representan firmas grandes que participan frecuentemente y que tienen una participación de mercado significativa. Las firmas tipo 0 ( $k = 0$ ), son las restantes que se tienen en la muestra. Los costos  $C_{ik}$  (costo privado de la firma  $i$  que es de tipo  $k$ ) se distribuyen según la función  $F_k(\cdot)$  (una función que es potencialmente diferente para cada tipo de firma) que tiene asociada una densidad  $f_k(\cdot)$  que se asume continua y estrictamente positiva. Un supuesto que se mantiene en este análisis es que la cantidad de jugadores participantes se determina de manera exógena, es decir en cada licitación la cantidad de jugadores actuales es la misma que la cantidad de jugadores potenciales y dado este supuesto no se debe modelar una etapa previa en donde las firmas deciden si participar o no. Cada firma  $i$  de tipo  $k$  decide la oferta (estrategia óptima)  $b_{ik}$  como solución a un problema de optimización en el que maximiza su beneficio esperado, balanceando el beneficio que obtiene si gana (que es creciente en su oferta) con la probabilidad de ganar (que es decreciente en su oferta). Formalmente el problema de optimización es el siguiente:

$$\begin{aligned} \max_b \Pi_k(b_i; c_i, X_\ell, N_\ell) &= \max_{b_i} (b_i - c_i) \prod_{j \in n_k \setminus \{i\}} (1 - F_k(s_k^{-1}(b_i) | X_\ell, N_\ell))^j \\ &\quad \prod_{j \in n_{k'}} (1 - F_{k'}(s_{k'}^{-1}(b) | X_\ell, N_\ell))^j \\ &= \max_{b_i} (b_i - c_i) \prod_{j \in n_k \setminus \{i\}} (1 - G_k(b_i | X_\ell, N_\ell))^j \prod_{j \in n_{k'}} (1 - G_{k'}(b_i | X_\ell, N_\ell))^j, \end{aligned}$$

donde  $k \neq k' \in \{0, 1\}$  y  $G_k(b | X_\ell, N_\ell) = F_k(s_k^{-1}(b) | X_\ell, N_\ell)$  es la probabilidad de que el jugador  $j \in n_k \setminus \{i\}$  de tipo  $k$  presente una oferta menor que  $b$  y lo mismo para un jugador de tipo  $k'$ . La variable  $X_\ell$  representa características propias

y exógenas de la licitación  $\ell$  (por ejemplo la estimación del gobierno sobre el costo de la obra en una licitación pública) y  $N_\ell$  es la cantidad de participantes en dicha licitación. La función  $s_k(\cdot)$  denota la estrategia de equilibrio de una firma de tipo  $k$ . Este tipo de juego asimétrico y su equilibrio (existencia y unicidad) ha sido caracterizado entre otros por Lebrun (1996), Lebrun (1999), Maskin y Riley (2000a), Maskin y Riley (2000b), Maskin y Riley (2003). La condición de primer orden del jugador  $i \in n_{k\ell}$  es:

$$(b_{ki} - c_{ki}) = \frac{1}{(n_{k\ell} - 1) \frac{g_k(b_{ki}|X_\ell, N_\ell)}{1 - G_k(b_{ki}|X_\ell, N_\ell)} + n_{k'} \frac{g_{k'}(b_{ki}|X_\ell, N_\ell)}{1 - G_{k'}(b_{ki}|X_\ell, N_\ell)}}.$$

Esta condición de primer orden junto con las condiciones de frontera  $s_k(\bar{c}) = \bar{b}$  para  $k = 0, 1$  caracterizan la solución del juego. Las primitivas del modelo son las funciones de distribución de los costos de cada tipo de participante,  $F_k(\cdot)$ , con  $k = 0, 1$ , que son funciones condicionales dadas las características específicas de cada subasta  $X_\ell$  y la cantidad de jugadores  $N_\ell$ . Las variables condicionantes se incluyen para controlar por el hecho de que las licitaciones no son homogéneas, es decir no es exactamente el mismo bien el que se licita en cada transacción.

A los efectos de lograr identificación de los costos privados (no observados) a partir de las ofertas (observadas) se asume lo siguiente.

1. En una licitación  $\ell$  participan  $n_\ell$  jugadores neutrales al riesgo, donde  $n_\ell$  es un elemento del conjunto  $[\underline{n}, \bar{n}]$  y  $n$  es un número mayor o igual a 2.
2. El vector  $(X_\ell, N_\ell)$  se distribuye según la función  $Q_m$  cuya densidad asociada es  $q_m$  para todo  $\ell = 1, \dots, L$ .
3. Para cada licitación  $\ell$  y para cada tipo  $k$ , las variables  $C_{k\ell}$  son iid (independientes e idénticamente distribuidas) con función de distribución acumulada  $F_k(\cdot)$  cuya densidad es  $f_k(\cdot)$ , ambas funciones son funciones condicionadas a las variables  $(X_\ell, N_\ell)$ .

4. Las ofertas de cada tipo de jugador  $B_k$  se distribuyen según la función  $G_k$  con densidad  $g_k$  para  $k = 0, 1$ . Estas funciones también deben entenderse como condicionadas a  $(X_\ell, N_\ell)$ .
5. Participación Exógena: Las funciones de distribución de los costos privados no dependen de la cantidad de oferentes, i.e.,  $F_k(c|X, N) = F_k(c|X)$ , para todo  $k = 0, 1$  y  $X$ .

Todos estos supuestos son estándares en la literatura de subastas con entrada exógena. Más aún este último supuesto no requiere que las variables  $(X_\ell, N_\ell)$  sean independientes entre sí. En otras palabras, el supuesto de entrada exógena es consistente con la presencia de correlación entre  $X$  y  $N$ . Esto permite racionalizar, por ejemplo, el hecho de que subastas más atractivas (contratos de obra pública más grandes, por ejemplo) atraen una mayor cantidad de participantes. Una vez descrito el modelo, el próximo paso es verificar que el mismo está identificado. El problema de identificación consiste en establecer si a partir de los datos observados (ofertas) pueden recuperarse las variables no observadas (costos privados) que son los elementos estructurales clave que se necesitan para realizar cualquier tipo de análisis, en particular los tests que se pretenden implementar. El argumento para identificar los costos privados (suprimiendo la dependencia con respecto a  $X$  y  $N$  y obviando los subíndices que indican asimetría) se basa en la siguiente sencilla observación. Dado que los costos privados y las ofertas observadas están vinculados a través de una función estrictamente creciente en equilibrio, la función  $s_k$ , se obtiene lo siguiente,

$$F(c) = Pr(C \leq c) = Pr(s(C) \leq s(c)) = Pr(B \leq s(c)) = G(s(c)).$$

La clave es la monotonicidad de la estrategia de equilibrio  $s$  que permite identificar a la función  $F$  a partir de la función  $G$ . Si bien la función  $G$  no es directamente observable, es fácilmente estimable a partir de las ofertas observadas. Este argumento que establece identificación lleva de manera

natural a un procedimiento de estimación no paramétrico en dos etapas como el propuesto por Guerre et al. (2000). En particular, se utilizan las  $n_k$  ofertas de tipo  $k$  para estimar  $G_k(c|X_\ell, N_\ell)$  y  $g_k(c|X_\ell, N_\ell)$  a través de métodos no paramétricos como *kernels* o polinomios locales. Luego, partir de las condiciones de primer orden del modelo que caracterizan las estrategias (ofertas) óptimas pueden obtenerse los costos privados. Estos costos dependen de las funciones de distribución de las ofertas, de las densidades de las mismas y de la cantidad de participantes. En particular, en este modelo, los costos privados se obtienen a partir de la siguiente ecuación

$$c_{ki} = b_{ki}^\ell - \frac{1}{(n_{k\ell} - 1) \frac{g_k(b_{ki}^\ell | X_\ell, N_\ell)}{1 - G_k(b_{ki}^\ell | X_\ell, N_\ell)} + n_{k'\ell} \frac{g_{k'}(b_{ki}^\ell | X_\ell, N_\ell)}{1 - G_{k'}(b_{ki}^\ell | X_\ell, N_\ell)}}. \quad (18)$$

Una vez obtenida una muestra de (pseudo) costos privados, Guerre et al. (2000) proponen implementar un estimador no paramétrico para obtener la función de densidad estimada  $f_k$ . De esta manera el procedimiento es enteramente no paramétrico. Otros autores han propuesto implementar esta segunda etapa a través de estimadores de Momentos Generalizados (GMM) lo que mejora las propiedades estadísticas del estimador resultante (ver Aryal et al., 2017).

La principal ventaja de este método indirecto es, por lo tanto, que no se requiere resolver la ecuación diferencial que caracteriza al equilibrio bayesiano de Nash, por lo tanto, no es necesaria una solución analítica del sistema de ecuaciones.

Con los costos recuperados en la etapa 1 (ecuación 18) se pueden implementar los tests (no paramétricos) para verificar si se rechaza o no la hipótesis nula de ausencia de colusión sin necesidad de implementar estrictamente la etapa 2 del procedimiento. Los costos recuperados a partir de este modelo se comparan luego con costos provenientes de un modelo de colusión que se describe a continuación.

## 6.2. El modelo con colusión (Modelo B)

El segundo modelo que se analiza es uno que toma en cuenta el hecho de que un subconjunto de firmas colude. Es necesario entonces describir la manera en la cual dicha colusión se lleva a cabo. Se asume que las firmas coluden de manera eficiente y en particular son las firmas de tipo 1 las que actúan de esta manera<sup>15</sup>. Este supuesto implica que el ambiente que se analiza es el más favorable posible para este tipo de prácticas (ver por ejemplo Marshall y Marx, 2007). La existencia de colusión puede generar asimetrías entre los participantes. Por lo tanto, es natural utilizar un modelo con jugadores asimétricos (esto es con al menos dos tipos de jugadores) en este contexto. Es importante aclarar que tanto el modelo de competencia como el modelo de colusión permiten que los jugadores sean ex-ante asimétricos. No obstante, las fuentes de asimetría en cada modelo son diferentes. En el Modelo A de competencia, las asimetrías reflejan que los participantes no pueden considerarse iguales, porque son más grandes, más eficientes, etc. En el Modelo B, más allá de que estos factores pueden estar presentes, el hecho de que existe un cártel operando en el mercado implica que dichos jugadores son diferentes al resto. En particular, se asume que el cártel puede controlar las ofertas de sus miembros y puede eliminar toda competencia proveniente del cártel, es decir que solamente existirá un único jugador “serio” en cada licitación que será a su vez la firma más eficiente, es decir aquella que puede efectuar la oferta más baja dentro del cártel debido a que es la que tiene los costos más bajos. En términos de la notación desarrollada esto implica que  $n_1 = 1$  (desde la perspectiva de los miembros del cártel, que se asume son jugadores de tipo 1). Como se mencionó anteriormente este es el escenario más propicio para la existencia de colusión y a los efectos de verificar

---

<sup>15</sup>Se puede asumir que solo un sub grupo dentro de las firmas de tipo 1 colude, con lo cual se estaría ante 3 tipos de jugadores, esto es lo que hacen Aryal y Gabrielli (2013). Para simplificar la exposición suponemos que todas las firmas de tipo 1 están incluidas en el cártel.

si los datos son consistentes con este modelo, no es necesario detallar las reglas exactas que utiliza el cártel para repartir los beneficios entre sus miembros. También se asume que los demás jugadores que no pertenecen al cártel no conocen la existencia del mismo. Los otros aspectos del modelo son los mismos que los del Modelo A. Cada miembro del cártel sabe que solamente compete con los jugadores de tipo 0, ya que asumimos que todos los jugadores de tipo 1 están en el cártel. La condición de primer orden para este tipo de jugador es entonces

$$1 = (b_{1i} - c_{1i}) \left[ \frac{n_0 f_1(s_0^{-1}(b_{1i}))}{[1 - F_0(s_0^{-1}(b_{1i}))] s_0'(s_0^{-1}(b_{1i}))} \right].$$

Esta ecuación junto a la correspondiente para jugadores tipo 0 determinan un nuevo sistema de ecuaciones diferenciales que, junto a las condiciones de frontera correspondientes, caracterizan un único equilibrio para las estrategias óptimas en un modelo que incorpora colusión. Notar que los jugadores de tipo 0 no cambian su comportamiento entre modelos. Solamente los jugadores de tipo 1 se comportan de manera diferente en cada caso. Utilizando un argumento de identificación similar al mencionado para el modelo de competencia se obtiene la siguiente ecuación que caracteriza los costos privados a partir de las ofertas y de las funciones de distribución y densidad de dichas ofertas.

$$\hat{c}_{1i} = \xi_{1i}(\cdot) = b_{1i} - \frac{1}{n_0 \ell \frac{g_0(b_{1i}|\cdot)}{1 - \hat{G}_0(b_{1i}|\cdot)}}, \quad (19)$$

donde se abusa de la notación y se utiliza  $b_{1i}$  para indicar a la oferta más baja entre las ofertas de los jugadores de tipo 1, ya que bajo el esquema colusivo las demás ofertas son meramente ofertas “fantasmas” y por lo tanto son arbitrarias, es decir no proveen información sobre los costos privados. Por lo tanto, para estimar el Modelo B se utilizará siempre la menor oferta entre los miembros del cártel. Al igual que con el modelo A, a los efectos de implementar los tests



no es necesario recuperar la función  $F_1(\cdot)$  de distribución de los costos de los jugadores de tipo 1. Los tests utilizados requieren como *inputs* a los costos privados (estimados), en este sentido, estrictamente se utiliza el primer paso de la metodología indirecta de estimación de Guerre et al. (2000).

### 6.3. Tests para detectar colusión

Como mencionamos anteriormente Aryal y Gabrielli (2013) plantean el problema de la “colusión” como un problema de testear por dominancia estocástica de primer orden, para lo cual utilizan el test de rangos y el test de Kolmogorov-Smirnov. A partir de cada modelo (competencia y colusión) se pueden obtener los costos privados de cada firma asociados a las ofertas observadas en los datos y a partir de estos costos se pueden implementar los tests mencionados. En particular, luego de la estimación de cada modelo se cuenta con los siguientes conjuntos de datos

$$M_A := \{X_\ell, n_{1\ell}, \{\hat{c}_{1i}^A\}_{i=1}^{n_{1\ell}}\}, \ell = 1, 2, \dots, L,$$

$$M_B := \{X_\ell, n_{1\ell}, \{\hat{c}_1^B\}\}, \ell = 1, 2, \dots, L,$$

donde  $\hat{c}_{1i}^j$  y  $\hat{c}_1^j$  son los costos recuperados a partir de las ecuaciones (18) y (19), respectivamente. Una importante ventaja de utilizar el enfoque estructural para analizar este tipo de problema es que permite no solamente establecer identificación de una manera transparente y por lo tanto se puede relacionar a los datos con los elementos estructurales sino que también provee condiciones necesarias para racionalizar el modelo. En este sentido, elegir entre dos modelos sería equivalente a verificar cuál de las dos condiciones impuestas (competencia o colusión) se observa en los datos. Dado que en este contexto

la única diferencia que existe entre los dos modelos son los costos privados de lo jugadores de tipo 1, es natural pensar que el criterio para diferenciar los modelos y seleccionar el apropiado debe solamente basarse en estos dos conjuntos de costos. Bajo los supuestos de Aryal y Gabrielli (2013), en particular el de entrada exógena al juego, y a partir de las ecuaciones (18) y (19) se advierte que si se observa la misma oferta  $b$  el costo implicado bajo colusión debe ser menor al que implica un modelo de competencia, i.e.,  $c_1^A \geq c_1^B$ . En otros términos, dado que el efecto de la colusión es deprimir la competencia en los mercados, si se observa la misma oferta en ambos esquemas, necesariamente debe darse la desigualdad mencionada entre los costos. Esto es consistente con el hecho que bajo colusión los márgenes de beneficios (diferencia entre el menor costo y la oferta realizada) son mayores en relación a los márgenes bajo competencia. Una implicación necesaria de este fenómeno es que la función de distribución de los costos recuperados bajo el Modelo A,  $M_A$ , debería dominar en sentido estocástico de primer orden a la distribución de los costos recuperados bajo el Modelo  $M_B$ .

Para testear por dominancia estocástica de primer orden se utiliza el test de rangos que se basa en estadísticos “U”. Para definir el ranking relativo de las variables aleatorias, sea  $c^A = \{c_{11}^A, c_{12}^A, \dots, c_{1n_\ell}^A\}$  y  $c^B = \{c_1^B\}$  para cada licitación  $\ell$ . A los efectos de utilizar este procedimiento hay que combinar y ordenar los costos estimados de ambos modelos. Definir  $U_i^\ell = 1$  si  $i$ ésima observación es de Modelo A,  $U_i^\ell = 0$  en otro caso. Luego el Test de Rangos de Mann-Whitney-Wilcoxon tiene la siguiente forma

$$U = \frac{1}{L} \sum_{\ell=1}^L \frac{1}{n_1^\ell} \sum_{i=1}^{2n_1^\ell} U_i^\ell.$$

Para implementar el test debemos evaluar  $\Pr(U \leq t)$ . En muestras grandes,  $U$  es aproximadamente Normal,

$$Z = \frac{U - \mu_u}{\sigma_u} \Rightarrow N(0, 1),$$

con  $\mu_u = \frac{n_1 n_2}{2}$  y  $\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$  si no hay empates.

El segundo test que se utiliza para discriminar entre colusión y competencia es el conocido Test de Kolmogorov-Smirnov cuyo estadístico se define,

$$KS_L = \sup_{c \in [\underline{c}, \bar{c}]} \left| F_1^A(c) - F_1^B(c) \right|.$$

En este caso tenemos que  $\Pr(KS_L \leq t) = \frac{\sqrt{2\pi}}{t} \sum_{k=1}^{\infty} e^{-(2k-1)^2 \pi^2 / (8t)^2}$ . No obstante, esta distribución no se puede usar porque  $c$  es estimado, no observado directamente. Dado que la distribución asintótica es compleja, una alternativa es utilizar *bootstrap*.

Cuadro 1. Fracción de decisiones correctas cuando hay Colusión

	Exponencial	Exponencial	Exponencial	Normal	Normal	Normal	Normal
Tipo 1	1.5	0.8	0.4	(1.5,1.5)	(1.5,0.9)	(0.9, 1.2)	(2.9,0.3)
Tipo 0	0.5	0.5	0.5	(3.5,0.8)	(2.5,0.3)	(2.8,0.4)	(2.5,0.3)
W-MW	1	1	1	1	1	1	1
KS	1	1	1	1	1	1	1
	Normal	Log Norm	Log Norm	Log Norm	Weibull	Weibull	Uniforme
Tipo 1	(1.9, 0.3)	(1.2, 1.3)	(2.0, 0.3)	(1.5, 1.0)	(3.0, 2.5)	(2.5,1.8)	(1.0, 4.0)
Tipo 0	(2.5, 0.3)	(2.9, 0.7)	(2.8, 0.2)	(2.5, 0.8)	(3.0, 4.0)	(2.5,3.0)	(1.0,4.0)
W-MW	1	1	1	1	1	1	1
KS	1	1	1	1	0.97	1	1

Nota: proporción de decisiones correctas. Valores p menores a 0.05. Cada test es aplicado 1000 veces para cada parámetro.

Una pregunta natural que surge en este contexto es si los tests utilizados efectivamente pueden detectar colusión cuando ésta existe en los datos. Para esto los autores desarrollaron un experimento Monte Carlo en donde el

verdadero proceso generador de datos contenía colusión. En particular se simulan ofertas de subastas con jugadores asimétricos en donde los jugadores tipo 1 coluden. El procedimiento utilizado es luego confiable si en este esquema se rechaza la hipótesis nula en favor de la alternativa. Los resultados obtenidos, porcentaje de rechazo de la hipótesis nula, confirman que el poder del test para detectar colusión es alto.

## **7. Conclusiones**

Este capítulo intenta describir los aspectos centrales y algunas de las contribuciones más significativas al Enfoque Estructural para analizar datos de subastas. Los aportes que se han desarrollado en los últimos 20 años son numerosos y sería imposible citarlos a todos. En particular nos hemos focalizado en el análisis de subastas o licitaciones de primer precio a sobre cerrado. Analizar este tipo de mecanismos no es restrictivo. La mayoría de las bases de datos de subastas disponibles para investigación empírica provienen de este formato. Más aún, otros formatos son estratégicamente equivalentes o posibles de ser analizados desde una perspectiva similar. En el enfoque estructural para el análisis de subastas el aspecto clave es el vínculo entre el modelo de teoría de juegos subyacente y el modelo econométrico a estimar. En particular, se supone que las ofertas observadas provienen del equilibrio bayesiano de Nash. La identificación (no paramétrica) del modelo le permite al investigador proponer un Modelo Econométrico Estructural. Laffont y Vuong (1996) desarrollan una serie de resultados que muestran en qué casos los modelos pueden identificarse a partir de las ofertas y la cantidad de jugadores y en qué casos esto no es posible. Al respecto el modelo más general que está identificado es el que considera valuaciones privadas

afiliadas. En el otro extremo, los modelos con valuaciones comunes no están generalmente identificados. Una vez que se establece la identificación del modelo, el siguiente paso es la estimación del mismo. Dentro del enfoque estructural se distinguen dos tipos de métodos de estimación: procedimientos directos e indirectos. Los métodos directos se desarrollaron primero en la literatura y abarcan procedimientos de estimación basados en máxima verosimilitud y procedimientos basados en simulaciones. Estos últimos son computacionalmente más atractivos ya que no necesitan que se resuelva de manera explícita el equilibrio del modelo, algo que puede llegar a ser altamente demandante en cálculos o en algunos casos imposible. Luego Guerre et al. (2000) introdujeron el uso de los métodos indirectos que descansan fundamentalmente en una observación muy simple pero crucial: existe una relación entre la distribución de las ofertas observadas y la de las valuaciones privadas a través de la función que describe a la estrategia de equilibrio que es estrictamente creciente. Esta observación permite la identificación (no paramétrica) de los elementos estructurales que caracterizan el modelo (i.e., la función de distribución de las valuaciones privadas) y lleva naturalmente a proponer un método de estimación que se implementa en dos etapas. Dada la naturaleza no paramétrica de la estrategia de identificación, los métodos indirectos típicamente utilizan estrategias de estimación no paramétricas. También, recientemente se han propuesto estrategias semi paramétricas que poseen mejores propiedades estadísticas. Finalmente, dado que la colusión (comportamientos cooperativos) es un problema omnipresente en los mercados de subastas que se traduce en una fuente importante de asimetrías es que se han desarrollado tests dentro de esta línea de análisis para detectar comportamientos colusivos.

## Referencias

- Aryal, G. y M. F. Gabrielli (2013). Testing for collusion in asymmetric first price auction. *International Journal of Industrial Organization*, 31, 26-35.
- Aryal, G., M. F. Gabrielli y Q. Vuong (2017): Semiparametric Estimation of First-Price Auction Models, <http://arxiv.org/abs/1407.7140>.
- Asker, J. (2008). A study of the internal organization of a bidding cartel. *American Economic Review*, 100(3), 724-762.
- Ausubel, L. M., y P. Milgrom (2006). The Lovely but Lonely Vickrey Auction, en P. Cramton, Y. Shoham, y T. Steinberg (eds.), *Combinatorial Auctions*, pp. 17–40. MIT Press.
- Bajari, P. (2001). Comparing competition and collusion: a numerical approach. *Economic Theory*, 18, 187–205.
- Campo, S., I. Perrigne y Q. Vuong (1998). Asymmetry and joint bidding in ocs wildcat auctions. Mimeo, University of Southern California.
- Comanor, W.S. y M.A. Schankerman (1976). Identical bids and cartel behavior. *Bell Journal of Economics*, 7, 281–286.
- Donald, S. y H. Paarsch (1993). Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. *International Economic Review*, 34:121–148.
- Donald, S. y H. Paarsch (1996). Identification, estimation, and testing in parametric empirical models of auctions within the independent private values paradigm. *Econometric Theory*, 12, 517-567.
- Fan, J. y I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall.
- Feinstein, J.S., M.K. Block y F.C. Nold (1985). Asymmetric information and collusive behavior in auction markets. *The American Economic Review*, 75(3), 441-460.

- Guerre, E., I. Perrigne y Q. Vuong (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68, 525-574.
- Haile, P., H. Hong y M. Shum (2003). Nonparametric tests for common values in first-price sealed-bid auctions. NBER Working Paper Series.
- Harrington, J.E. (2008). Detecting Cartels, en Buccirosi, P. (ed.), *Handbook in Antitrust Economics*. MIT Press.
- Harsanyi, J. (1967). Games with incomplete information played by bayesian players. *Management Science*, 14, 159-82, 320-34, 486-502.
- Hendricks, K., J. Pinkse y R. Porter (2003). Empirical implications of equilibrium bidding in first-price, symmetric, common value auctions. *Review of Economic Studies*, 70, 115-145.
- Ibragimov, I. y R. Has'minskii (1981). *Statistical Estimation. Asymptotic Theory*. Springer-Verlag New York, Inc.
- Laffont, J. J. y Q. Vuong (1996). Structural analysis of auction data. *American Economic Review*, 86, 414-420.
- Laffont, J.J., H. Ossard y Q. Vuong (1995). Econometrics of first-price auctions. *Econometrica*, 63, 953-980.
- Lang, K. y R.W. Rosenthal (1991). The contractor's game. *The RAND Journal of Economics*, 22, 329-338.
- Lebrun, B. (1996). Existence of an equilibrium in first price auctions. *Economic Theory*, 7, 421-443.
- Lebrun, B. (1999) First-price auction in the asymmetric N bidder case. *International Economic Review*, 40, 125-142.
- Li, T. y Q. Vuong (1997). Using all bids in parametric estimation of first-price auctions. *Economic Letters*, 55, 321-325.
- Marshall, R., M. Meurer, J. Richard, J. y W. Stromquist (1994). Numerical analysis of asymmetric first price auctions. *Games and Economic Behavior*, 7, 193-220.

- Marshall, R. y L.M. Marx (2007). Bidder collusion. *Journal of Economic Theory*, 133, 374-402.
- Marshall, R. y M. Meurer (2001). The Economics of Auctions and Bidder Collusion, en Chatterjee, K., Samuelson, W. (eds.), *Game Theory and Business Applications*. Kluwer Academic Publishers, Norwell.
- Maskin, E. y J. Riley (2000a). Asymmetric auctions. *Review of Economic Studies*, 67, 413-438.
- Maskin, E. y J. Riley (2000b). Equilibrium in sealed high bid auctions. *Review of Economic Studies*, 67, 439-454.
- Maskin, E. y J. Riley (2003) Uniqueness of equilibrium in sealed high-bid auctions. *Games and Economic Behavior*, 45, 395-409.
- McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57, 995-1026.
- Milgrom, P. y R. Weber (1982). A Theory of Auctions and Competitive Bidding. *Econometrica*, 50, 1089-1122.
- Paarsch, H. (1992). Deciding between the common and private value paradigms in empirical models of auctions. *Journal of Econometrics*, 51, 191- 215.
- Pakes, A. y D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57, 1027-1057.
- Pesendorfer, M. (2000). A study of collusion in first-price auctions. *Review of Economic Studies*, 67, 381-411.
- Porter, R. H. (1995). The role of information in U.S. offshore oil and gas lease auctions. *Econometrica*, 63(1), 1-27.
- Porter, R. y D. Zona (1993). Detection of bid-rigging in procurement auctions. *Journal of Political Economy*, 101, 518-538.
- Porter, R. y D. Zona (1999). Ohio School Milk Markets: an analysis of bidding. *The RAND Journal of Economics*, 30, 263-288.



- Riley, J. y W. Samuelson (1981). Optimal auctions. *American Economic Review*, 71, 381–392.
- Rothkopf, M. (1969). A model of rational competitive bidding. *Management Science*, 15(7), 362-373.
- Vickrey, W. (1961). Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance*, 16(1), 8-37.
- Wilson, R. (1969). Competitive bidding with disparate information. *Management Science*, 15(7), 446-448.
- Wilson, R. (1977). A bidding model of perfect competition. *Review of Economic Studies*, 44(3), 511-518.



## Capítulo 4. Fundamentos de econometría espacial aplicada

Marcos Herrera\*

### 1. Introducción

La econometría espacial es una rama econométrica dedicada al análisis de *efectos espaciales* en modelos de regresión en datos de corte transversal y de panel. Anselin (2010) ubica el nacimiento del área en el año 1979 producto de una serie de publicaciones que sentaron las primeras bases metodológicas (Bartels y Ketellapper, 1979; Bennett, 1979; Hordijk, 1979; Paelinck y Klaassen, 1979; Cliff y Ord, 1981). Prácticamente una década más tarde, la publicación del libro *Spatial econometrics: Methods and models* (Anselin, 1988) consolida las principales líneas metodológicas que marcaron la primera agenda de investigación dentro del área. Desde los inicios, el principal método de estimación de los modelos espaciales fue por máxima

---

\* Agradezco a Jesús Mur por ser mi principal mentor en estos temas. Errores u omisiones son de exclusiva responsabilidad del autor.

verosimilitud (*ML, Maximum Likelihood*). Sin embargo, los aportes realizados por Kelejian y Prucha (1998, 1999) y Lee (2003) impulsaron la estimación por el variables instrumentales (*IV, Instrumental Variables*) y método de momentos generalizados (*GMM, Generalized Method of Moments*). Adicionalmente, el libro de LeSage y Pace (2009) resume los aportes por estimación bayesiana.

Tradicionalmente, los modelos que incorporaban explícitamente el espacio (topología geográfica) estaban confinados a temas de ciencia regional, economía geográfica o urbana. Tal es así que la primera vez que se empleó el término *econometría espacial* fue para brindar fundamentación metodológica a modelos regionales y urbanos (Paelinck, *Annual Meeting of the Dutch Statistical Association*, Tilburg, 1974). Sin embargo, el actual entendimiento del espacio no está restringido a cuestiones geográficas ampliando el uso de las técnicas al espacio económico y social (distancia socio-económica). La diversidad de aplicaciones que utilizan herramientas de econometría espacial incluye temas de redes sociales, criminología, finanzas públicas, ciencia política, psicología, salud, educación, historia, transporte, marketing, entre muchos otros.

En este capítulo se revisan y organizan los desarrollos más consolidados de econometría espacial enfatizando un rango de tópicos relativos a la especificación e interpretación de los efectos espaciales. Estos efectos pueden dividirse en: (1) *dependencia o interacción espacial*, generalmente entendida por su versión más débil de *autocorrelación espacial* y (2) *heterogeneidad espacial* que es un tipo de inestabilidad estructural proveniente del término de error (heteroscedasticidad) o de los coeficientes de las variables del modelo. Entre estos efectos, la heterogeneidad espacial puede ser resuelta mediante herramientas econométricas estándares. En cambio, la dependencia espacial es el efecto que mayor dificultad presenta y por eso nos dedicaremos casi exclusivamente a su tratamiento. Debido a la amplia literatura sobre el tema,

el interés se centrará en modelos con variable dependiente continua que buscan incorporar dependencia espacial.

Dentro del área económica, la econometría espacial adquirió mayor relevancia a partir de la década del 90 por diferentes desarrollos teóricos que involucraron cuestiones geográficas y de interacción social (por ejemplo, Krugman, 1991; Manski, 1993; Brock y Durlauf, 1995, entre otros avances). Puede afirmarse que esta demora obedece a cuestiones históricas, la teoría económica minimizó la influencia del espacio o lo supuso neutral al comportamiento de los mercados y a los procesos productivos y distributivos de bienes. Un simple repaso de los principales libros de texto utilizados en las carreras de ciencias económicas permite comprobar que las principales corrientes teóricas no mencionan casi ningún aspecto de la localización de los individuos o de los mercados. Dicho olvido se traslada directamente a la enseñanza de la econometría. El análisis econométrico de datos con información georreferenciada no está presente en los libros de econometría de base y el problema de dependencia es relegado casi con exclusividad a series temporales. Este último punto impulsa el desarrollo del presente capítulo con la intención de cubrir una temática prácticamente ausente en cursos de econometría intermedia y avanzados.

El resto del capítulo está organizado de la siguiente manera. En la sección 2 se presentan los principales elementos de econometría espacial. La sección 3 desarrolla los principales modelos espaciales en corte transversal y en datos de panel. En la sección 4 se discute la interpretación de los modelos que incluyen efectos espaciales endógenos. La sección 5 ilustra los desarrollos teóricos mediante dos ejemplos empíricos. Finalmente, los comentarios finales son incluidos en la sección 6.

## 2. Conceptos y elementos espaciales

Esta sección contiene conceptos y temas relacionados a la dependencia espacial en su versión más débil de autocorrelación espacial. Antes de explorar la autocorrelación espacial, se presenta una discusión sobre los datos espaciales y sus particularidades. Este conocimiento nos permitirá hablar de autocorrelación espacial de una forma más precisa y entender la necesidad de la matriz de contactos espaciales, elemento clave en el área.

### 2.1. Sobre la naturaleza de los datos

El primer aspecto a destacar, y no siempre mencionado, es sobre la naturaleza de los datos espaciales. Existen diferentes tipos de datos espaciales y esto conlleva desarrollos estadísticos particulares para los mismos.

Cressie (1993) realiza una taxonomía de datos según los supuestos del proceso estocástico espacial (*spatial random field*). Sin una excesiva formalización, este proceso estocástico puede definirse como un conjunto de variables aleatorias  $\mathbf{y}$ , indexadas por localización  $i$ ,  $\{y_i, i \in D\}$ , donde  $D$  es un índice proveniente de un subconjunto del espacio euclidiano ( $D \subset \mathbb{R}^2$ ). Dependiendo del supuesto realizado sobre  $D$ , los datos pueden clasificarse como geoestadísticos (*geostatistical data*), areales o regionales (*lattice data*) y de patrón de puntos (*point pattern data*).

Los datos geoestadísticos son datos provenientes de una superficie continua ocupando un subconjunto fijo  $D$ . Por ejemplo,  $D$  puede ser el área de la Argentina y los datos representan la superficie de precipitaciones. Los datos regionales provienen de un índice  $D$  fijo, siendo una colección finita, regular o

irregular, de puntos o áreas que particionan al espacio euclidiano. Ejemplo de este tipo es el área de la Argentina particionada con datos del producto bruto geográfico por provincias. En el caso del patrón de puntos son datos en donde  $D$  es un proceso puntual en  $\mathbb{R}^2$ . La diferencia de este último tipo de datos respecto a los anteriores proviene del hecho de que los puntos en  $D$  son el resultado de un proceso aleatorio, siendo el principal interés la localización de un evento en particular. Este último tipo de dato espacial es rara vez analizado en nuestra área. Para mayor detalle sobre estos tipos de datos una excelente referencia es Cressie (1993).

Las diferencias existentes entre los datos geoestadísticos y regionales pueden relacionarse con la noción de modelos de datos (*data models*) que son abstracciones de la realidad en forma manejable y almacenada por medio de los sistemas de información geográfica (*GIS, Geographic Information Systems*). El modelo raster almacena campos (mallas o cuadrículas regulares) pertenecientes a distribuciones espaciales continuas y representadas como superficies. En el modelo vectorial se almacenan objetos discretos representados por líneas, puntos y polígonos. En economía, estos objetos pueden corresponder a jurisdicciones, viviendas, radios censales, etc. Consecuentemente, los datos regionales y el modelo vectorial se encuentran asociados de una forma natural y son típicamente analizados en econometría espacial.

Sin importar el tipo de dato espacial a analizar, existen problemas que son comunes a todos. Uno de ellos es el problema de unidad de área modificable (*Modifiable Areal Unit Problem, MAUP*), versión geográfica de la falacia ecológica, en donde las conclusiones basadas en una agregación particular de las áreas o regiones pueden cambiar si se agregan los mismos datos en un conjunto diferentes de áreas o regiones (Waller y Gotway, 2004). El efecto de agregaciones geográficas diferentes fue investigado por Openshaw y Taylor (1979), con datos electorales de 99 *counties* en Iowa, incorporando el término *MAUP* para describir el problema. Estos autores construyeron

todos los posibles agrupamientos en grandes distritos alcanzando diferentes e inexplicables correlaciones con rangos de valores desde -0.97 hasta +0.99. El problema tiene dos componentes distinguibles: uno es el efecto escala o agregación y el otro es el efecto agrupamiento o zonificación (Openshaw y Taylor, 1979; Openshaw, 1984). El primer efecto radica en la obtención de distintas inferencias estadísticas cuando el mismo conjunto de datos es agregado en diferentes resoluciones espaciales, especialmente cuando se agregan pequeñas áreas en grandes regiones (por ejemplo, datos de familias agregados para radios censales o para comunas de la Ciudad Autónoma de Buenos Aires). El segundo efecto se refiere a la variación en el resultado debido al agrupamiento alternativo de las regiones en la misma o similar escala espacial (por ejemplo, datos de familias agregados por comunas o por distritos escolares de CABA). Ambos efectos pueden generar errores que afectarán la validez de los resultados. El problema, igual que en el caso de la falacia ecológica, no tiene solución y esto implica la necesaria cautela que debe hacerse al extraer conclusiones inferenciales, más aún si son base para la toma de decisiones políticas (para mayor detalle puede consultarse Arbia, 2012).

Otro problema relacionado a la naturaleza de los datos espaciales, o debido al proceso de medición, ocurre al integrar diferentes tipos de datos en una única base de información. Por ejemplo, es común que la toda información no esté disponible al nivel de escala deseada: imaginemos que deseamos investigar el impacto de la delincuencia o contaminación ambiental en un análisis hedónico de precio de vivienda. Los datos de ubicación de las viviendas son una nube de puntos (modelo vectorial) pero los datos de contaminación es una superficie (campo continuo, modelo raster) o en el caso del índice de delincuencia puede estar construido a nivel barrial (polígonos, modelo vectorial). La necesaria transformación de los datos espaciales a la escala de interés genera el problema de cambio de soporte (*COSP, Change Of Support Problem*). El término “soporte” hace referencia al tamaño y volumen de cada



base pero también incluye la forma, el tamaño y la orientación espacial de los objetos o campos representados. El paso de un soporte involucra generar nuevas variables relacionadas a las originales pero con diferencias espaciales y estadísticas. Existen alternativas para atenuar el problema aunque no lo resuelven completamente. Para conocer el detalle de estas alternativas puede consultarse el artículo de Gotway y Young (2002) y referencias allí indicadas.

Una vez revisada la naturaleza de los datos espaciales discutiremos la definición de dependencia en estos datos y el enfoque más utilizado para introducir este elemento en la especificación econométrica.

## 2.2. Definiendo autocorrelación espacial

Es típico en econometría establecer independencia de las observaciones permitiendo así un tratamiento estadístico manejable. Una implicación de este supuesto es que los valores observados  $y_i$ , para un individuo  $i$ , son estadísticamente independientes de otros valores  $y_j$ , para el individuo  $j$ ,  $E(y_i y_j) = E(y_i) E(y_j) = 0$  donde  $E(\cdot)$  es el operador de esperanza, suponiendo datos centrados en media. Sin embargo, cuando los individuos representan a periodos temporales o a regiones en el espacio este supuesto es demasiado restrictivo y deberá relajarse permitiendo alguna forma de dependencia.

La definición más habitual de dependencia espacial es por medio de la autocorrelación espacial que puede definirse, vagamente, como la similitud de valores en localizaciones próximas. La autocorrelación puede ser positiva, en el sentido que un valor alto (bajo) de una variable aleatoria en una posición espacial es acompañado por un vecindario de valores altos (bajos). También es posible una autocorrelación espacial negativa con un valor elevado (bajo) en una posición geográfica y valores bajos (elevados) en el vecindario. Ya

sea si la autocorrelación espacial es positiva o negativa, su existencia implica que una muestra de datos contiene menor información que una muestra no correlacionada. Esta pérdida de información debe ser explícitamente considerada en la estimación y este es el problema esencial en temas de econometría aplicada con datos espaciales.

De manera más formal, la autocorrelación espacial es una versión re-escalada de la siguiente condición de momentos (covarianzas):

$$\text{Cov}(y_i, y_j) = E(y_i y_j) - E(y_i) E(y_j) \neq 0, \quad \forall i \neq j, \quad (1)$$

donde  $y_i, y_j$  son observaciones de una variable aleatoria en la localización  $i$  y  $j$  en el espacio, es decir, todo par  $(i, j)$  posee información geográfica específica medida por latitud y longitud.

Dada una muestra de  $n$  observaciones en un corte transversal, existen  $(n^2 - n) / 2$  covarianzas a ser estimadas, considerando simetría. Dicha estimación es imposible de realizar y no puede resolverse incrementando el tamaño muestral, siempre habrá más covarianzas que datos disponibles. Este es un problema de parámetros incidentales y para resolverlo se necesita imponer restricciones sobre la forma en que las observaciones interactúan entre sí.

Existen al menos dos enfoques utilizados para estimar la covarianza espacial. En geoestadística, todos los pares de localizaciones son clasificados según la distancia que los separa entre sí y la covarianza es expresada como una función continua de dicha distancia, comúnmente conocida como variograma o semi-variograma. Esta alternativa supone que el proceso estocástico es continuo en el espacio (datos geoestadísticos). La segunda alternativa supone que el proceso espacial es discreto (datos regionales) construyendo una medida de similitud para la variable aleatoria y otra medida de similitud para la localización. La similitud de la localización se materializa en la construcción del vecindario para cada observación bajo una matriz de pesos espaciales. Esta alternativa es la elección mayoritaria en econometría espacial y será desarrollada a continuación.

### 2.3. Importancia de la matriz de pesos espaciales

El enfoque preponderante en econometría espacial utiliza la matriz de pesos espaciales para brindar una medida de similitud de las localizaciones. Esta matriz comúnmente se denominada  $\mathbf{W}$  (es similar a la matriz  $\mathbf{G}$  en redes sociales) y describe la conectividad entre  $n$  unidades que se encuentran localizadas en un espacio bidimensional. La construcción de  $\mathbf{W}$  está basada en al menos dos supuestos clave sobre la estructura espacial: (1) un criterio de conectividad que define qué unidades pueden ser consideradas vecinas entre si, y (2) un supuesto de pesos espaciales que operacionaliza cómo los vecinos se afectan entre si.

Para entender cómo la matriz  $\mathbf{W}$  permite simplificar el problema de dependencia espacial formalizaremos la discusión. Supóngase un proceso autorregresivo espacial *SAR* (*Spatial AutoRegresive*) tal que la variable de estudio  $\mathbf{y}$  se distribuye espacialmente en tres regiones de la siguiente forma:

$$\begin{aligned} y_i &= \alpha_{ij}y_j + \alpha_{ik}y_k + u_i, \\ y_j &= \alpha_{ji}y_i + \alpha_{jk}y_k + u_j, \\ y_k &= \alpha_{ki}y_i + \alpha_{kj}y_j + u_k, \\ u_i; u_j; u_k &\sim i.i.d. (0; \sigma^2). \end{aligned} \tag{2}$$

es decir, en la primera ecuación, el valor  $y$  en la región  $i$  depende del valor de  $y$  en la región  $j$  y  $k$ , más un término aleatorio que se distribuye de manera idéntica e independiente entre las localizaciones. Lo mismo puede decirse de las regiones  $j$  y  $k$ .

En términos matriciales, la expresión (2) se reduce a:

$$\mathbf{y} = \mathbf{A}\mathbf{y} + \mathbf{u}, \tag{3}$$

donde  $\mathbf{y}' = [y_i \ y_j \ y_k]$ ,  $\mathbf{u}' = [u_i \ u_j \ u_k]$  con la matriz  $\mathbf{A}$  conteniendo los siguientes elementos:

$$\mathbf{A} = \begin{bmatrix} 0 & \alpha_{ij} & \alpha_{ik} \\ \alpha_{ji} & 0 & \alpha_{jk} \\ \alpha_{ki} & \alpha_{kj} & 0 \end{bmatrix}.$$

El problema con el sistema (3) es que, bajo un corte transversal, hay más parámetros que observaciones (3 observaciones y 7 parámetros). Puede imponerse simetría, reduciendo el número de parámetros pero, de todas formas, no podrá estimarse. Por lo tanto, es necesario buscar una solución alternativa para poder estimar la dependencia subyacente en los datos.

La solución radica en imponer un conjunto de restricciones sobre las relaciones de dependencia. De esta forma, la estructura de la matriz  $\mathbf{A}$  es reparametrizada de la siguiente manera:

$$\mathbf{A} = \rho \mathbf{W},$$

donde  $\rho$  es un parámetro a estimar que captura el efecto promedio de interacción entre todas las regiones y  $\mathbf{W}$  es la matriz de pesos espaciales. Los elementos de  $\mathbf{W}$  son:

$$\mathbf{W} = \begin{bmatrix} 0 & w_{ij} & w_{ik} \\ w_{ji} & 0 & w_{jk} \\ w_{ki} & w_{kj} & 0 \end{bmatrix}, \quad (4)$$

tal que se ha intercambiado a los parámetros  $\alpha$ 's (parámetros originales del modelo) por  $w_{ij}$  (coeficientes exógenos al modelo) y el modelo queda expresado como  $\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{u}$ , donde  $\mathbf{W} \mathbf{y}$  es interpretado como el rezago espacial de  $\mathbf{y}$  por analogía a series temporales<sup>1</sup>.

---

<sup>1</sup>En series temporales, un proceso autorregresivo de orden 1, AR(1):  $y_t = \alpha y_{t-1} + u_t$ , puede representarse como  $\mathbf{y} = \alpha \mathbf{L} \mathbf{y} + \mathbf{u}$ , siendo  $\mathbf{L}$  el operador de rezagos temporales. La diferencia entre  $\mathbf{L}$  y  $\mathbf{W}$  radica en la multidireccionalidad espacial y esto genera que las técnicas estadísticas temporales no tengan las propiedades deseadas en el espacio.

Generalizando, la matriz  $\mathbf{W}$  será de orden  $n \times n$ , siendo  $n$  el tamaño muestral. Cada elemento de  $\mathbf{W}$  es denominado peso espacial,  $w_{ij}$ . Los pesos espaciales capturan la vecindad, siendo diferentes de cero cuando las regiones  $i$  y  $j$  son consideradas que interactúan entre sí. Por convención, ninguna región puede interactuar consigo misma, dando como resultado que la diagonal principal de  $\mathbf{W}$  contenga todos sus elementos iguales a cero,  $w_{ii} = 0$ .

La matriz de pesos espaciales ocupa una posición clave y su elección condicionará al resto del análisis. El problema central radica en cómo se construyen los pesos de la matriz  $\mathbf{W}$ . En la práctica econométrica, la matriz de contactos es construida mediante diferentes criterios. Estos criterios van desde el uso de la posición geográfica hasta el uso de flujos que capturan interacciones sociales y utilizan otras fuentes de información socio-económica.

La forma más tradicional para construir la matriz es bajo criterios geográficos, siguiendo la primera ley de la geografía (Tobler, 1970): “Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes.” Entre los criterios geográficos, puede definirse vecindad por contigüidad, por alguna función de distancia, por  $k$  vecinos más cercanos o alguna combinación de estas opciones.

El criterio contigüidad considera como vecinos a los polígonos que comparten límites. Hay varias opciones dentro de este criterio: tipo torre (vecinos al norte-sur y este-oeste) o tipo reina (todos los polígonos que comparten límite), nombres alusivos a los movimientos del ajedrez. Cuando la información es de tipo puntual (como el caso de viviendas), la aplicación de este criterio puede realizarse mediante la creación regiones artificiales (polígonos Voronoi o Thiessen).

Una opción más natural para la información de tipo puntual es utilizar algún criterio de distancia. Es posible construir diversas medidas, por ejemplo: vecinos son aquellas unidades que se encuentran a una distancia inferior a un determinado umbral, definido por un radio  $d^*$ . También es posible usar este

criterio cuando se cuenta con datos de áreas, en este caso la distancia entre polígonos se calculará utilizando los centroides de los polígonos. Dependiendo de la aplicación empírica, es posible cambiar el centroide por otro punto más relevante del polígono. Por ejemplo, si se trabaja con información de países puede usarse la ubicación de cada ciudad capital.

El problema con los criterios de contigüidad y de distancia por umbral es que pueden dar lugar a regiones aisladas que no tienen vecinos. Esto puede suceder debido a que la densidad de los puntos no es regular en el mapa o, bajo contigüidad, cuando se trabaja con mapas que contienen islas que naturalmente no comparten límites con otras regiones.

En el caso de contigüidad, el problema de regiones sin vecinos puede salvarse asignando manualmente vecinos a estas unidades. Bajo el criterio de distancia por umbral, puede determinarse un umbral cuya amplitud asegure que cada observación contenga al menos un vecino. Este criterio es conocido como max-min, el umbral será la máxima distancia del vecino más cercano de todas las unidades. Un umbral elevado puede ocasionar que existan regiones con una excesiva cantidad de vecinos dando lugar a un gran número de conexiones irrelevantes.

Otro criterio de uso habitual es el de  $k$  vecinos más cercanos. Utilizando latitud y longitud, se va eligiendo como vecino al punto más cercano hasta obtener el número de vecinos establecidos,  $k$ . Bajo este criterio todas las regiones poseerán la misma cantidad de vecinos evitando el problema de unidades aisladas o unidades con excesiva cantidad de vecinos.

Una vez definido el criterio de vecindad, resta operacionalizar cómo los individuos se afectan entre sí. Una forma simple es utilizando una elección binaria, con  $w_{ij} = 1$  cuando  $i$  y  $j$  cumplen con el criterio de vecindad elegido y  $w_{ij} = 0$  cuando no. Si el criterio utilizado es distancia, los pesos binarios pueden reemplazarse por alguna función de distancia entre unidades espaciales. Por ejemplo, utilizando la función la inversa  $w_{ij} = f(d_{ij}) = 1/d_{ij}$ , tal que a mayor

distancia menor relación entre  $i$  y  $j$ , cumpliendo de esta forma con la ley de Tobler. La literatura aplicada también ha combinado los criterios de distancia o contigüidad con otra información geográfica como el perímetro del polígono, la extensión de la frontera en común, y otras características geográficas.

Otras especificaciones de pesos espaciales combinan información socio-económica y con información geográfica. Continuando con el ejemplo de distancia podría definirse cada peso como  $w_{ij} = [|z_i - z_j| + 1]^{-1} \times 1/d_{ij}$ , donde  $z$  es una variable que captura una distancia socio-económica (PBI, flujo comercial, etc.). En otras aplicaciones, directamente se aplica un concepto de distancia socio-económica:  $w_{ij} = [|z_i - z_j| + 1]^{-1}$  con  $z$  siendo el ingreso per cápita o el porcentaje de población con alguna característica (tasa de pobreza, porcentaje de algún grupo étnico). El problema con la distancia socio-económica es que la exogeneidad del peso es dudosa en numerosas ocasiones.

Definidos los pesos espaciales, lo habitual es trabajar con alguna transformación de la matriz debido a que mejora las propiedades estadísticas de los estimadores. La transformación más utilizada es la normalización por fila, en donde los nuevos pesos se obtienen como  $w_{ij}^* = w_{ij}/\sum_j w_{ij}$ , tal que la suma de cada fila de  $W$  será igual a la unidad:  $\sum_j w_{ij}^* = 1$ . Esta transformación tiene implicancias a la hora de interpretar el coeficiente espacial  $\rho$ : medirá el impacto promedio de los vecinos sobre una región. Otra opción es normalizar por columna,  $w_{ij}^* = w_{ij}/\sum_i w_{ij}$ , en cuyo caso el coeficiente espacial capturará el impacto promedio de la región sobre los vecinos.

Respecto al orden de vecindad, puede ser de interés conocer el impacto de los vecinos de los vecinos. Por ejemplo, bajo el criterio de distancia por umbral, el vecindario de segundo orden estará compuesto por las regiones que se encuentren a una distancia  $d^* < d_{ij} \leq 2d^*$ . Para estos casos, los ordenes superiores a 1 se identifican con un supra-índice:  $\mathbf{W}^{(j)}$ ,  $\forall j \geq 2$ , siendo  $j$  el orden de vecindad.

Por lo general la elección del criterio de vecindad es un a priori del investigador. Aunque existen un conjunto de lineamientos útiles para la elección de la matriz establecidos por Griffith (1996): (i) Es mejor utilizar una especificación razonable de la matriz de pesos geográficos que considerar todas las conexiones nulas; (ii) modelos con bajos ordenes deben ser preferidos por sobre los modelos de altos ordenes; (iii) en general, es mejor emplear una matriz de pesos sub-identificada que una sobre-identificada.

Sin importar cuál sea el caso empírico, el desarrollo econométrico tradicional supone que  $\mathbf{W}$  contiene elementos exógenos a la especificación econométrica y la elección de los vecinos no debería realizarse mediante variables consideradas en el modelo analizado.

### 3. Especificaciones espaciales alternativas

La creación de la matriz de pesos espaciales brinda la oportunidad de contrastar e incluir nuevas variables al modelo econométrico. A nivel empírico, la inclusión de estos efectos espaciales ha sido dirigida por dos tipos de enfoques. Por un lado, los modelos espaciales pueden especificarse siguiendo una teoría sobre la interacción entre los agentes, enfoque conocido como *theory-driven*. Hay numerosas investigaciones que utilizan teorías bien establecidas e incorporan externalidades en forma de *spillovers* espaciales. En contraposición a este tipo de enfoque, gran parte de la literatura aplicada ha sido típicamente dirigida por problemas que no tienen un sustento teórico formal. Este enfoque, denominado *data-driven*, encuentra justificación por problemas ya mencionados: los datos no se encuentran disponibles en la misma escala y localización al proceso espacial de interés, existen errores de medición u omisión de variables que generan dependencia espacial. El enfoque *data-driven*



busca obtener estimadores consistentes, insesgados y eficientes controlando por la estructura espacial presente en los datos. En esta sección exploraremos las alternativas de especificación que han sido utilizadas bajo ambos enfoques.

### 3.1. Modelos espaciales de corte transversal

Un punto de partida muy utilizado en investigaciones empíricas es un modelo no-espacial, usualmente estimado por mínimos cuadrados ordinarios (*MCO*). Este modelo lineal no-espacial es un modelo lineal general con la siguiente especificación:

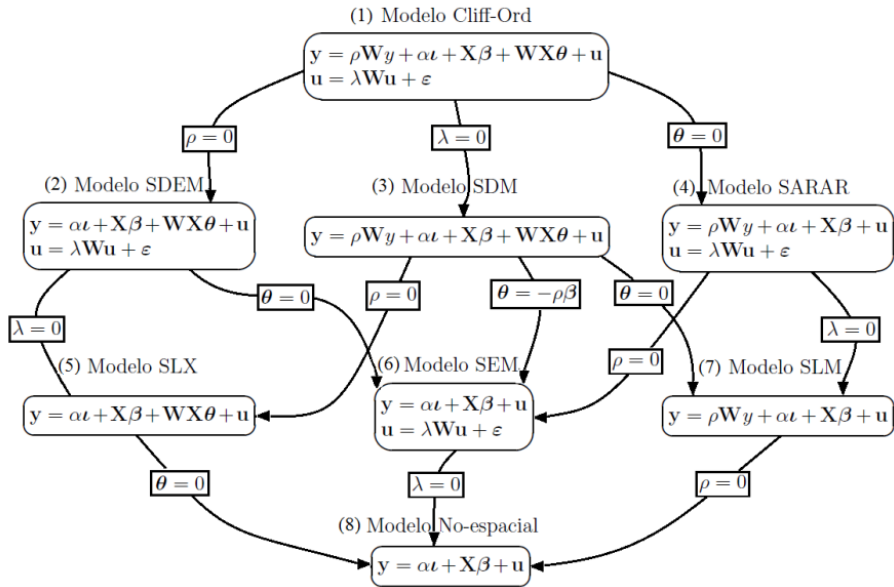
$$\mathbf{y} = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (5)$$

siendo la variable dependiente  $\mathbf{y}$  un vector de dimensión  $(n \times 1)$ ,  $\mathbf{1}$  es un vector  $(n \times 1)$  de unos asociados al término constante  $\alpha$ ,  $\mathbf{X}$  es una matriz de variables explicativas de orden  $(n \times K)$ ,  $\boldsymbol{\beta}$  es un vector de parámetros desconocidos de orden  $(K \times 1)$  y  $\mathbf{u} = (u_1, u_2, \dots, u_n)'$  es un vector de términos aleatorios de dimensión  $(n \times 1)$ , donde  $u_i$  es independiente e idénticamente distribuido para todo  $i$  con media cero y varianza constante  $\sigma^2$ . Se supone que el modelo cumple con los supuestos habituales en la parte sistemática.

La detección de efectos espaciales en la ecuación (5) permite introducir efectos espaciales de diversas formas, tal como puede observarse en la Figura 1. Un recorrido habitual es introducir efectos espaciales secuencialmente usando como punto de partida el modelo lineal, estrategia “de lo particular a lo general” (*Stge, Specific-to-general*). Para aplicar esta estrategia puede utilizarse una batería de tests estadísticos de multiplicadores de Lagrange, *LM*, simples y robustos, que permiten detectar las posibles estructuras espaciales omitidas (consúltese Anselin et al., 1996, para mayor detalle). En muchas

aplicaciones, previo al desarrollo de la estrategia *Stge*, se realiza un análisis exploratorio de datos espaciales (mapas, detección de agrupamientos, test espaciales univariados, detección de *outliers* espaciales, etc.) que permite establecer conjeturas sobre la importancia del espacio, generalmente bajo el enfoque *data-driven*. Anselin (2005) muestra cómo realizar dicho análisis con el programa GeoDa, Bivand (2010) ofrece similares alternativas bajo R y Herrera (2015) presenta las opciones disponibles en Stata.

Figura 1: Modelos espaciales alternativos en corte transversal



Fuente: elaboración propia siguiendo modelos propuestos en Elhorst (2014b).

Otra estrategia de especificación, una vez detectada la importancia del espacio, es comenzar por un modelo espacial lo más complejo posible,

estrategia “de lo general a la particular” (*Gets, General-to-specific*) (Hendry, 1979). En la literatura espacial el modelo más complejo es conocido como modelo general de Cliff-Ord (modelo 1, Figura 1):

$$\begin{aligned} \mathbf{y} &= \rho \mathbf{W}\mathbf{y} + \alpha\iota + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \\ \mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \varepsilon, \end{aligned} \quad (6)$$

donde  $\mathbf{W}\mathbf{y}$  representa el efecto de interacción espacial endógeno o efecto espacial sustantivo,  $\mathbf{W}\mathbf{X}$  son los efectos de interacción exógenos y  $\mathbf{W}\mathbf{u}$  es el efecto de interacción espacial en el término de error o efecto espacial residual. El parámetro  $\rho$  es denominado coeficiente de rezago espacial,  $\lambda$  es el coeficiente de autocorrelación espacial de los errores y  $\boldsymbol{\theta}$  son los coeficientes espaciales de dependencia local.

La ecuación (6) posee una estructura similar al conocido modelo de Manski (1993) de efectos sociales. En el modelo de Manski las interacciones sociales son especificadas como el resultado esperado de un equilibrio social. La similitud entre ambos modelos ha sido destacada por Lee (2004, 2007) entre otros. La especificación del modelo es:

$$\begin{aligned} \mathbf{y} &= \rho_m E[\mathbf{y}|\mathbf{a}] + \alpha\iota + \mathbf{X}\boldsymbol{\beta} + E[\mathbf{X}|\mathbf{a}]\boldsymbol{\theta} + \mathbf{u}, \\ \mathbf{u} &= \lambda_m E[\mathbf{u}|\mathbf{a}] + \varepsilon, \end{aligned} \quad (7)$$

donde  $\mathbf{a}$  indexa unidades no-solapadas de vecinos. En la literatura de redes sociales,  $\rho_m$  captura el efecto endógeno del resultado esperado de los vecinos sobre la variable dependiente,  $\boldsymbol{\theta}$  captura el efecto contextual o exógeno de las características grupales y  $\lambda_m$  captura el efecto correlacionado de características no observables. Luego de algunas manipulaciones, la forma reducida del modelo puede escribirse como:

$$\mathbf{y} = \alpha\iota + \mathbf{X}\boldsymbol{\beta} + E[\mathbf{X}|\mathbf{a}] \frac{(\rho_m \boldsymbol{\beta} + \boldsymbol{\theta})}{1 - \rho_m} + \frac{\rho_m}{1 - \rho_m} E[\mathbf{u}|\mathbf{a}] + \mathbf{u}, \quad (8)$$

y esta última ecuación permite visualizar el problema de identificación planteado por Manski, “problema de reflexión”: no es posible separar el

efecto endógeno del efecto exógeno a menos que se impongan restricciones adicionales. Los modelos espaciales se diferencian de esta literatura en el sentido que no se busca estimar el efecto “vecindario” ya que cada unidad observada se encuentra conectada por una estructura que se supone conocida por medio de la especificación de  $\mathbf{W}$ . Salvando esta diferencia, el modelo general de Cliff-Ord presenta similares problemas de identificación imposibilitando su uso en la estrategia *Gets*. Es decir, el modelo espacial más complejo solo puede incorporar dos efectos espaciales de los tres posibles ( $\rho$ ,  $\lambda$  o  $\theta$ ) necesitando imponer restricciones en el efecto restante.

Las especificaciones alternativas según la restricción sobre el modelo de Cliff-Ord se presentan en la Figura 1, modelos (2), (3) y (4). Si se impone la restricción  $\theta = 0$ , el modelo resultante es denominado *SARAR* (*Spatial AutoRegressive with AutoRegressive error*), con dependencia espacial y sustantiva, impulsado principalmente por Kelejian y Prucha (1998) en sus desarrollos teóricos. Este modelo es frecuentemente utilizado en teoría econométrica ya que contiene los dos términos espaciales más complejos de estimar. Sin embargo, a nivel empírico ha sido escasamente utilizado ya que proporciona cierta protección contra la falta de especificación en una parte del modelo pero contamina las estimaciones de otras partes (Pace y Zhu, 2012).

LeSage y Pace (2009) proponen al modelo espacial de Durbin (*SDM, Spatial Durbin Model*) para iniciar la especificación general (modelo 3, Figura 1):

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}. \quad (9)$$

Los motivos que llevan a esta elección son comprensibles: (1) la omisión de un término relevante en el error solo genera pérdida de eficiencia; pero (2) los costos de ignorar efectos espaciales en la variable dependiente y/o la omisión de variables explicativas espaciales relevantes generan estimadores sesgados e inconsistente de los coeficientes restantes. Elhorst (2014b) reconoce la ventaja del modelo *SDM*, sin embargo, no descarta al modelo *SDEM* (*Spatial Durbin*

*Error Model*) como una especificación de interés y competitiva en términos predictivos (modelo 2, Figura 1). Los análisis comparativos entre los modelos *SDM* y *SDEM* muestran que son prácticamente indistinguibles cuando no existen motivos teóricos para la incorporación de un rezago espacial endógeno.

Es interesante mencionar que el modelo *SDM* anida la mayor cantidad de modelos: si  $\theta = \mathbf{0}$  el modelo se reduce a un *SLM* (*Spatial Lag Model*), si  $\rho = 0$  el modelo alcanzado es el *SLX* (*Spatial Lag in X's*), y mediante una combinación no-lineal ( $\theta = -\rho\beta$ ) el *SDM* puede reducirse a un modelo *SEM* (*Spatial Error Model*). Este último resultado se alcanza reemplazando  $\theta = -\rho\beta$  en la ecuación (9):

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \alpha \mathbf{1} + \mathbf{X}\beta + \mathbf{W}\mathbf{X}(-\rho\beta) + \mathbf{u}.$$

Reagrupando términos en la última ecuación y reemplazando la notación  $\rho$  por  $\lambda$  se obtiene el modelo *SEM*:

$$\mathbf{y} = \alpha^* \mathbf{1} + \mathbf{X}\beta + (\mathbf{I} - \lambda \mathbf{W})^{-1} \boldsymbol{\varepsilon}.$$

Debido a la dificultad de estimar modelo más complejos, en la primera etapa de desarrollo de econometría espacial los modelos insignia eran el *SLM* y el *SEM*. El modelo de rezago espacial tuvo una presencia preponderante en esta etapa inicial por dos razones. Por una parte, bajo un esquema económico de teoría de juegos entre agentes, con funciones de pago cuadráticas en sus variables de elección, el *SLM* puede ser visto como la condición de primer orden o función de reacción que surge del problema de optimización simultáneo de  $n$  individuos (Brueckner, 2003). Otra razón estaba justificada por cuestiones de estimación, la inclusión de un término espacial endógeno evita la inconsistencia del estimador y, además, el modelo de rezago espacial obtiene una estimación del *spillover* espacial que es muy atractiva a nivel empírico. El otro modelo modelo con buena presencia en los inicios fue el modelo *SEM* que permite una ganancia de eficiencia respecto al modelo no-espacial. Una crítica empírica al modelo *SEM* sucede cuando la autocorrelación espacial es débil:

los resultados tienden a ser similares a los obtenidos por una estimación no-espacial robusta. Aunque en términos inferenciales, ignorar la presencia de un *SEM* tiende a subestimar la varianza del modelo.

El modelo espacial más simple en términos de estimación es el *SLX*:

$$\mathbf{y} = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \quad (10)$$

donde  $\boldsymbol{\theta}$  captura el efecto espacial de las variables explicativas de los vecinos a cada observación. Estos efectos son conocidos como contextuales en la literatura de modelos sociales y tienen un impacto local<sup>2</sup>.

Dentro de la literatura, el *SLX* no ha sido muy utilizado en aplicaciones empíricas y menos aún en trabajos teóricos. La falta de interés en econometría teórica proviene del hecho que no ofrece dificultad en su estimación, el modelo puede estimarse por *MCO*. Pero en cuestiones aplicadas Halleck y Elhorst (2015) aconsejan no descartar los resultados que ofrece este modelo.

### 3.2. Modelos espaciales en datos de panel

La taxonomía de los modelos presentados en la Figura 1 puede extenderse fácilmente a modelos espacio-temporales estáticos. Para pasar de información de corte transversal a información de datos de panel solo es necesario incorporar un sub-índice  $t$ , que va de  $t = 1, \dots, T$ , apilando las observaciones del corte transversal periodo a periodo. Sin modificar ningún supuesto en el

---

<sup>2</sup>Este modelo es muy utilizado en análisis de efectos pares o contextuales. Por ejemplo, la relación entre rendimiento académico ( $\mathbf{r}$ ) y nivel socio-económico ( $\mathbf{nes}$ ) individual y contextual puede especificarse como  $\mathbf{r} = \iota\alpha + \mathbf{nes}\beta_1 + \overline{\mathbf{nes}}\beta_2 + \mathbf{u}$ , en donde  $\overline{\mathbf{nes}} = \mathbf{W} \times \mathbf{nes}$  es el efecto contextual o rezago espacial del nivel socio-económico, siendo  $\mathbf{W}$  construida en función de pertenencia a la misma escuela.

término de error, este tipo de extensión espacio-temporal da lugar a los modelos *pooled*. Por ejemplo, el modelo *pooled* de Cliff-Ord puede escribirse como:

$$\begin{aligned} \mathbf{y}_t &= \rho \mathbf{W}\mathbf{y}_t + \alpha t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \mathbf{u}_t, \\ \mathbf{u}_t &= \lambda \mathbf{W}\mathbf{u}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \dots, T, \end{aligned} \quad (11)$$

donde las unidades de corte transversal ( $i = 1, \dots, n$ ) están apiladas por periodo temporal  $t$  y el término  $\boldsymbol{\varepsilon}_t$  es un vector de dimensión  $n$  de errores aleatorios *i.i.d.*

En econometría espacial se trabaja habitualmente con paneles micro, es decir, con un número amplio de unidades de corte transversal y un número reducido de periodos temporales. Por otra parte, las técnicas espaciales requieren que los datos de panel se encuentren balanceados en el sentido que las mismas unidades espaciales deben ser observadas en cada uno de los periodos. Esto plantea un problema adicional ya que los datos de panel tienden a sufrir el fenómeno de desgaste muestral, es decir, pérdida información entre diferentes periodos, ocasionando paneles desbalanceados.

Las críticas a los modelos *pooled* es que omiten posibles fuentes de heterogeneidad. Por ejemplo, no consideran particularidades de cada región que pueden ser específicas e invariantes en el tiempo. Características como religión, normas jurídicas o estructura laboral no muestran variabilidad temporal y su omisión aumenta el riesgo de sesgo en las estimaciones. Una posible solución es incorporar en los modelos un término  $\boldsymbol{\mu}$ , de dimensión  $n$ , que capture el efecto específico individual de las regiones. Otra fuente de heterogeneidad puede provenir de particularidades específicas entre periodos temporales, como un año de crisis o recesión. Estos eventos causarán diferencias entre periodos temporales que posiblemente sesgarán las estimaciones. Como solución a esto puede añadirse un término  $\phi_t$  que controle por efectos temporales específicos.

Al incorporar los efectos específicos, individuales y temporales, el modelo en 11 es re-expresado como:

$$\begin{aligned} \mathbf{y}_t &= \rho \mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \mu + \phi_t \iota_n + \mathbf{u}_t, \\ \mathbf{u}_t &= \lambda \mathbf{W}\mathbf{u}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \dots, T. \end{aligned} \quad (12)$$

Los efectos específicos pueden ser considerados como efectos fijos o aleatorios. En el caso un modelo de efectos fijos, una variable ficticia es introducida por cada unidad espacial (sin considerar a la unidad base) y otra variable variable ficticia es utilizada para capturar el efecto temporal específico.

En el caso de los modelos con efectos aleatorios, debe especificarse la distribución de estos efectos, como por ejemplo:  $\mu \sim \mathcal{N}(0, \sigma_\mu^2 \mathbf{I}_n)$ ,  $\phi \sim \mathcal{N}(0, \sigma_\phi^2 \mathbf{I}_t)$ . Entonces, la especificación en el caso de considerar un *SDM* con efectos aleatorios será:

$$\begin{aligned} \mathbf{y}_t &= \rho \mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \underbrace{\mu + \phi_t \iota_n + \boldsymbol{\varepsilon}_t}_{\mathbf{u}_t}, \\ \boldsymbol{\varepsilon}_t &\sim \mathcal{N}[0, \sigma_\varepsilon^2 \mathbf{I}_n], \quad \mu \sim \mathcal{N}[0, \sigma_\mu^2 \mathbf{I}_n], \quad \phi_t \sim \mathcal{N}(0, \sigma_\phi^2). \end{aligned}$$

En el caso de considerar un *SDEM* con efectos aleatorios, la especificación vendrá dada por:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \underbrace{\mu + \phi_t \iota_n + \boldsymbol{\varepsilon}_t}_{\mathbf{u}_t}, \\ \boldsymbol{\varepsilon}_t &= \lambda \mathbf{W}\boldsymbol{\varepsilon}_t + \boldsymbol{\eta}_t, \\ \boldsymbol{\eta}_t &\sim \mathcal{N}[0, \sigma_\eta^2 \mathbf{I}_n], \quad \mu \sim \mathcal{N}[0, \sigma_\mu^2 \mathbf{I}_n], \quad \phi_t \sim \mathcal{N}(0, \sigma_\phi^2). \end{aligned}$$

Estas versiones aleatorias suponen que el término remanente de error  $\boldsymbol{\varepsilon}_t$  posee estructura espacial. Otras especificaciones modifican este supuesto tal que el mismo proceso espacial aplica tanto a los efectos individuales como al error remanente (ver Baltagi et al., 2003, y para una especificación alternativa a Kapoor et al., 2007).

Respecto a los estadísticos para detectar dependencia espacial existen diversos desarrollos. La presencia de estructura espacial residual y sustantiva



en modelos *pooled* puede ser detectada usando las versiones de los tests *LM* desarrolladas por Anselin et al. (2008) adaptadas a la estructura espacio-temporal. También existen versiones de los tests *LM* para modelos no-espaciales de efectos fijos (Debarsy y Ertur, 2010) y aleatorios (Baltagi et al., 2003).

Alternativamente, pueden utilizarse otros tests generales que permiten detectar la presencia de dependencia espacial. Recuérdese que, bajo un corte transversal, la dependencia espacial ha sido reducida a la estimación de un parámetro debido a la insuficiente información para estimar la correlación entre pares de observaciones. Con datos de panel, esta limitación es reducida dependiendo de la longitud de la dimensión temporal y es posible obtener estimaciones de la correlación entre las observaciones del corte transversal. Un estadístico que detecta dependencia de corte transversal es el CD de Pesaran (2004). Este estadístico puede ser aplicado a una amplia variedad de paneles, incluyendo dinámicos y no-estacionarios, ya sea bajo efectos fijos o aleatorios.

La literatura rápidamente ha ampliado el análisis de paneles estáticos hacia paneles dinámicos. La incorporación de dinámica temporal genera un modelo básico, que en su forma vectorial puede escribirse, para un periodo  $t$ :

$$\mathbf{y}_t = \tau \mathbf{y}_{t-1} + \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\mu} + \phi_t \iota_n + \boldsymbol{\varepsilon}_t, \quad (13)$$

con  $\boldsymbol{\varepsilon}_t \sim i.i.d(0, \sigma_\varepsilon^2)$ , con  $\mathbf{y}_t$  e  $\mathbf{y}_{t-1}$  de orden  $n \times 1$  y efectos específicos individuales  $\boldsymbol{\mu}' = [\mu_1, \mu_2, \dots, \mu_n]$  y temporales  $\phi_t$ , con  $t = 1, \dots, T$ .

La ecuación (13) permite capturar la persistencia temporal del proceso, incluyendo la heterogeneidad entre las regiones y entre periodos. Igual que en los modelos estáticos, los efectos específicos pueden ser tratados como efectos fijos o aleatorios.

Tal como sucedía en los modelos estáticos, los efectos espaciales pueden introducirse en la ecuación (13) en diferentes formas. Anselin et al. (2008) propone un conjunto de especificaciones:

- Modelo simultáneo espacio-temporal:

$$\mathbf{y}_t = \tau\mathbf{y}_{t-1} + \rho\mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mu + \phi_t\iota_n + \varepsilon_t, \quad (14)$$

- Modelo recursivo espacial puro:

$$\mathbf{y}_t = \gamma\mathbf{W}\mathbf{y}_{t-1} + \mathbf{X}_t\boldsymbol{\beta} + \mu + \phi_t\iota_n + \varepsilon_t, \quad (15)$$

- Modelo recursivo espacio-temporal:

$$\mathbf{y}_t = \tau\mathbf{y}_{t-1} + \gamma\mathbf{W}\mathbf{y}_{t-1} + \mathbf{X}_t\boldsymbol{\beta} + \mu + \phi_t\iota_n + \varepsilon_t, \quad (16)$$

- Modelo dinámico espacio-temporal:

$$\mathbf{y}_t = \tau\mathbf{y}_{t-1} + \gamma\mathbf{W}\mathbf{y}_{t-1} + \rho\mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mu + \phi_t\iota_n + \varepsilon_t, \quad (17)$$

El modelo dinámico espacio-temporal incorpora todas las alternativas de rezagos espacio-temporales de la variable dependiente. Las características de este modelo fueron analizadas por Yu et al. (2008). El modelo recursivo espacio-temporal, ecuación (16), ha sido aplicado en cuestiones de predicciones espacio-temporales (por ejemplo, véase Giacomini y Granger, 2004).

Los modelos previamente mencionados son tipos de modelos espacio-temporales *SLM* (*dynSLM*) ya que solo incorporan elementos dinámicos espacio-temporales endógenos. Esta clasificación puede ampliarse hacia modelos espacio-temporales de Durbin (*dynSDM*) en diferentes versiones:

$$\mathbf{y}_t = \tau\mathbf{y}_{t-1} + \rho\mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \mu + \phi_t\iota_n + \varepsilon_t, \quad (18)$$

$$\mathbf{y}_t = \gamma\mathbf{W}\mathbf{y}_{t-1} + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \mu + \phi_t\iota_n + \varepsilon_t, \quad (19)$$

$$\mathbf{y}_t = \tau\mathbf{y}_{t-1} + \gamma\mathbf{W}\mathbf{y}_{t-1} + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \mu + \phi_t\iota_n + \varepsilon_t, \quad (20)$$

$$\mathbf{y}_t = \tau\mathbf{y}_{t-1} + \gamma\mathbf{W}\mathbf{y}_{t-1} + \rho\mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \mu + \phi_t\iota_n + \varepsilon_t, \quad (21)$$

en donde se han introducido efectos espaciales locales,  $\mathbf{W}\mathbf{X}_t$ , además de los efectos ya presentes en los modelos previos.

Revisiones relativamente recientes de estos modelos pueden encontrarse en Anselin et al. (2008), Lee y Yu (2010), Elhorst (2010b) y Lee y Yu (2014).

### 3.3. Cuestiones inferenciales y asintóticas

La estimación de los modelos espaciales no puede realizarse por métodos simples. La inclusión de variables  $\mathbf{W}y$  o  $\mathbf{W}u$  provocan que el estimador *MCO* sea inconsistente y/o ineficiente, dependiendo del caso (Anselin, 1988, pp. 57-59). Como consecuencia, la literatura ha desarrollado métodos alternativos de estimación que incluyen opciones paramétricas, no paramétricas y semi paramétricas. Aquí solo se mencionarán las alternativas más empleadas como la estimación por máxima verosimilitud, *ML*. Otra alternativa de gran desarrollo en los 90 estuvo basada en variables instrumentales o método de momentos generalizados (*IV/GMM*), y también puede mencionarse la estimación bayesiana (LeSage y Pace, 2009, cap. 5).

La estimación de estos modelos presenta una serie de dificultades relacionadas a los parámetros espaciales y al comportamiento de las matrices  $\mathbf{W}$ ,  $(\mathbf{I} - \rho\mathbf{W})^{-1}$  y/o  $(\mathbf{I} - \lambda\mathbf{W})^{-1}$  presentes en todos los métodos de estimación. Una de ellas es la necesidad de imponer alguna restricción sobre el rango de valores que puede alcanzar  $\rho$ , lo mismo se aplica para  $\lambda$ . Sin restricciones sobre este parámetro, el espacio paramétrico da lugar a discontinuidades y no asegura la invertibilidad de la matriz  $(\mathbf{I} - \rho\mathbf{W})$ . Para evitar este problema y también facilitar la estimación de  $\rho$ , Ord (1981) propuso restringir los posibles valores del parámetro en un rango  $1/r_{min} < \rho < 1/r_{max}$ , donde  $r_{min}$  y  $r_{max}$  son los autovalores, mínimo y máximo, de la matriz  $\mathbf{W}$ . Bajo una matriz espacial estandarizada por filas automáticamente se impone una restricción sobre el máximo autovalor tal que  $\rho$  se encontrará restringido entre los valores  $1/r_{min} <$

$\rho < 1$ . Algunos autores como Kelejian y Robinson (1995) han sugerido limitar el espacio paramétrico de  $\rho$  al intervalo  $(-1, 1)$  en similitud al tratamiento que se realiza en series temporales, sin embargo este rango de valores puede ser demasiado restrictivo para el caso espacial.

En el caso de los modelos dinámicos espacio-temporales también deben cumplirse con ciertas restricciones de estacionariedad. Como Elhorst (2014b) explica, la estimación de un modelo dinámico que incorpore los términos  $\mathbf{y}_{t-1}$ ,  $\mathbf{W}\mathbf{y}_{t-1}$ ,  $\mathbf{W}\mathbf{y}_t$  será estable si se cumplen las condiciones:  $|\tau| < 1$  y  $|\tau| + |\gamma| + |\rho| < 1$ . En situaciones en donde esta condición no se cumpla, el modelo estará espacialmente cointegrado y una transformación del modelo será necesaria para su estimación (Yu et al., 2012).

En el caso de la estimación por *ML*, una dificultad adicional se encuentra en la maximización de la función de log-verosimilitud que involucra al determinante de una matriz de dimensión  $n \times n$ ,  $\ln|\mathbf{I} - \rho\mathbf{W}|$ . Este determinante debe ser evaluado repetidamente para cada valor estimado de  $\rho$  y el procedimiento puede ser problemático para grandes bases de datos. En la era de *big data*, donde puede tenerse millones de observaciones la precisión de este determinante es clave. Una solución propuesta por Ord (1975) es estimar dicho término mediante la siguiente transformación:

$$\ln|\mathbf{I} - \rho\mathbf{W}| = \sum_{i=1}^n \ln(1 - \rho r_i), \quad (22)$$

siendo  $r_i$  los autovalores de la matriz  $\mathbf{W}$ , que reduce la complejidad pero puede ser problemática en tiempo computacional para grandes bases de datos.

Respecto a la estimación por *IV/GMM*, los principales avances de este método fueron realizados por Kelejian y Prucha (1998, 1999), con generalizaciones realizadas por Kelejian y Prucha (2010), Arraiz et al. (2010) y Drukker et al. (2013). Las condiciones de momentos no incluyen la estimación del determinante, ecuación (22), estableciendo ventajas computacionales claras y notorias si nos encontramos con grandes bases de datos. Otra virtud de

este método es la posibilidad de instrumentalizar otras variables explicativas endógenas, además de la endogeneidad espacial. Este es un punto en donde *ML* o estimación bayesiana no presentan ninguna solución.

Una potencial desventaja de la estimación por *IV/GMM* es que no impone ninguna restricción sobre el espacio paramétrico en el proceso de estimación. Por lo tanto puede suceder que el valor estimado del coeficiente espacial ( $\rho$  o  $\lambda$ ) se encuentre fuera del rango de estabilidad dando lugar a estimaciones no interpretables.

En el caso de la estimación de los paneles espaciales dinámicos (estacionarios), hay diferentes alternativas que extienden los procedimientos de máxima verosimilitud condicional y no-condicional, así como existen procedimientos alternativos por *IV/GMM*. Elhorst (2010a) muestra diferentes procedimientos basados en *ML* y Yu et al. (2008) desarrollan un estimador de cuasi-máxima verosimilitud que muestra propiedades comparables a los estimadores *ML*.

Respecto a las propiedades asintóticas de los estimadores, en econometría estándar suponer que  $n \rightarrow \infty$  no incluye ningún tipo de información adicional. Pero en el espacio, el incremento del tamaño muestral necesita especificar de qué forma es expandido el tamaño muestral en el espacio geográfico. Elhorst (2014b, p. 55) comenta que existen dos tipos de asintótica comúnmente utilizadas en el contexto espacial. Una de ellas es la estructura “infill”, donde las observaciones espaciales se encuentran delimitadas en el espacio geográfico a medida que  $n \rightarrow \infty$ . Otra estructura es “increasing domain”, donde los límites geográficos de las regiones se expanden cuando  $n \rightarrow \infty$ . Por otra parte, hay dos tipos de diseño muestral en el espacio: (a) diseño estocástico donde las unidades espaciales son aleatoriamente seleccionadas; y (b) diseño fijo donde las unidades espaciales se encuentran relacionadas a un espacio no aleatorio, posiblemente irregular.

La literatura de econometría espacial se enfoca principalmente en la estructura asintótica “increasing domain” bajo un diseño muestral fijo esto es debido a que mayoritariamente se utilizan datos areales. Por lo tanto, una posición con mayor respaldo ha sido la estimación por efectos fijos, dado que el espacio es usualmente muestreado en forma exhaustiva y la inferencia realizada puede considerarse como poblacional. De todas formas, hay argumentos para realizar estimaciones por efectos aleatorios ya que no todos los estudios utilizan de forma exhaustiva el espacio geográfico y, por otra parte, los datos de panel para un intervalo temporal pueden considerarse como una muestra representativa de una superpoblación (concepto relacionado a un proceso generador de datos espacio-temporal, ver Graubard y Korn, 2002, y referencias allí citadas para mayor detalle).

Las propiedades de los estimadores dependen crucialmente de determinadas condiciones de regularidad que debe cumplir la matriz de pesos espaciales. Un requisito mencionado por Kelejian y Prucha (1998) establece que la sumatoria de las columnas y filas de las matrices  $\mathbf{W}$ ,  $(I - \rho\mathbf{W})^{-1}$ , antes de normalizar  $\mathbf{W}$ , debe estar uniformemente limitada en valor absoluto cuando  $n$  tiende a infinito. Además, Lee (2004) establece que la sumatoria de filas y columnas de  $\mathbf{W}$ , antes de estandarizar, no debe divergir a infinito a una velocidad igual o mayor que la tasa a la que crece el tamaño muestral  $n$ . El cumplimiento de al menos una de las anteriores condiciones permiten que la correlación entre dos unidades espaciales tienda a cero cuando la distancia que las separa tiende a infinito y de esta forma limitan la correlación a un rango de valores manejables. Por ejemplo, si se trabaja con el criterio contigüidad la primera condición es satisfecha automáticamente. En cambio, si el criterio utilizado es la inversa de la distancia la primera condición no es satisfecha pero si satisface la condición que establece Lee (2004).

Las anteriores condiciones establecen situaciones límites a la conectividad de la matriz de pesos. Pero aún cuando se cumplan, las propiedades finitas de

los estimadores pueden verse afectadas. Smith (2009) muestra que matrices espaciales fuertemente conectadas (medidas por el número medio de vecinos), con una densidad media de vínculos de 10% o superior, generan un sesgo descendente en los parámetros espaciales bajo *ML*. En todos los casos, este sesgo implica que la autocorrelación espacial positiva puede no ser detectada debido a la presencia de matrices densamente conectadas y el estimador *ML* tiende sistemáticamente a cero cuando  $n$  crece. Por lo general, la conectividad de la matriz por criterios geográficos son poco densas, no superando el mencionado umbral. El problema puede ser importante cuando la matriz es construida por criterios de redes sociales o no-espacial.

#### **4. Interpretación de los modelos con rezago espacial endógeno**

Un punto de interés, válido para los modelos de corte transversal y de datos de panel, es la interpretación de los resultados obtenidos. Por ejemplo, en los inicios de la especialidad el modelo *SLM* fue adoptado como un modelo central debido a que brinda una estimación de la interdependencia espacial. La interpretación de los resultados obtenidos era directa, en el sentido que un valor significativo del parámetro  $\rho$  era evidencia de la importancia de los efectos desbordamiento o contagio. Sin embargo, LeSage y Pace (2009) destacan la errónea interpretación de estos resultados y muestran la necesidad de considerar derivadas parciales del impacto de un cambio en una unidad de las variables explicativas.

Para ver más claramente el punto destacado por LeSage y Pace (2009), debe tenerse presente que los efectos desbordamiento en un modelo *SDM*, por ejemplo, se deben a la presencia de **Wy** (efecto global) y **WX** (efecto local):

$$\mathbf{y}_t = \rho \mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_t\boldsymbol{\theta} + \mathbf{u}_t. \quad (23)$$

Supongamos que se desea analizar el impacto del incremento unitario de la  $k$  –ésima variable explicativa, este incremento es considerado igual para todas las localizaciones. Bajo un modelo de regresión no espacial, el efecto total sobre la variable dependiente será igual a la estimación del coeficiente  $\beta_k$ , *ceteris paribus*, sin importar la localización. Pero ante un modelo como el presentado en (23), el efecto total depende de las unidades vecinas de cada localización y de la magnitud de los coeficientes que acompañan a las variables espaciales.

El efecto total puede ser descompuesto en un efecto directo y otro indirecto. El efecto directo es el impacto del cambio de la variable explicativa sobre la variable dependiente en cada localidad, este efecto tenderá a ser similar al obtenido por un modelo de regresión no espacial si  $\rho$  y  $\theta_k$  son cercanos a 0. El efecto indirecto se debe a la dinámica espacial generada por la presencia de  $\rho$  y  $\theta_k$  que afectará a todas las unidades el modelo.

Bajo un modelo de Durbin pueden distinguirse dos tipos de efectos indirectos producto de la interdependencia entre las unidades. Uno es de tipo global, afectando a todas las unidades:  $y_i$  aumenta inicialmente  $\beta_{ik}$  unidades cuando la explicativa aumenta en una unidad, y esto genera un nuevo incremento producto del cambio en las  $y'_j$ s de los  $S$  vecinos, capturado por el término endógeno  $\rho \mathbf{W}\mathbf{y}$ . Es decir, para cada  $i$  –ésima localidad existirá un nuevo impacto igual a  $\rho \sum_{j \neq i}^S w_{ij} \beta_{jk}$  unidades, retro-alimentando el cambio en la variable dependiente y que solo cesará al converger (suponiendo  $|\rho| < 1$ ). Este efecto es identificado como un *efecto espacial global* para distinguirlo del *efecto espacial local* que surge del incremento unitario en la  $k$  –ésima variable explicativa de los  $S$  vecinos de la localidad  $i$ :  $\sum_{j \neq i}^S w_{ij} \theta_{jk}$ . Este efecto espacial refuerza al efecto inicial  $\beta_k$  y es de tipo local en el sentido que no posee un efecto dinamizador como el generado por la presencia de  $\rho$ .



En términos matriciales, y de forma más compacta, el *efecto marginal total* en el modelo *SDM* puede ser representado por la matriz de derivadas parciales de  $y$  respecto al cambio en una unidad de la  $k$  –ésima variable explicativa en  $X$ , dado un periodo  $t$ :

$$\begin{aligned} \left[ \frac{\partial E(\mathbf{y})}{\partial x_{1k}} \quad \dots \quad \frac{\partial E(\mathbf{y})}{\partial x_{nk}} \right]_t &= (\mathbf{I}_n - \rho \mathbf{W})_t^{-1} \begin{bmatrix} \beta_k & w_{12}\theta_k & \dots & w_{1n}\theta_k \\ w_{21}\theta_k & \beta_k & \dots & w_{2n}\theta_k \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1}\theta_k & w_{n2}\theta_k & \dots & \beta_k \end{bmatrix}_t, \\ &= \mathbf{S}_t [\beta_k \mathbf{I}_n + \theta_k \mathbf{W}]_t, \end{aligned} \quad (24)$$

donde  $w_{ij}$  es el elemento  $(i, j)$  –ésimo de  $W$ ,  $\beta_k$  es el  $k$  –ésimo elemento del vector  $\beta$ ,  $\theta_k$  es el elemento  $k$  –ésimo del vector  $\theta$  y  $\mathbf{S}_t$  es igual a  $(\mathbf{I}_n - \rho \mathbf{W})_t^{-1}$ . Obsérvese que para la estimación de los efectos marginales se utiliza el valor esperado de la forma reducida del modelo,  $E(\mathbf{y}_t) = (\mathbf{I} - \rho \mathbf{W})_t^{-1} [\mathbf{X}_t \beta + \mathbf{W} \mathbf{X}_t \theta]$ .

En la expresión (24) puede observarse la presencia de un efecto directo y otro indirecto. El *efecto directo* es representado por los elementos de la diagonal principal de la ecuación (24). Este cambio será diferente para cada región por lo que no existe un único efecto directo. Por su parte, el *efecto indirecto (spatial spillover)*, proviene de los elementos fuera de la diagonal principal de (24) y, nuevamente, será diferente para cada región. En este caso se tiene un sumatorio de efectos para cada localidad.

Para el caso de un *SLM*,  $\mathbf{y}_t = \rho \mathbf{W} \mathbf{y}_t + \mathbf{X}_t \beta + \varepsilon_t$ , los *efectos marginales totales* se simplifican ya que  $\theta_k = 0$  siendo la expresión final igual a  $\beta_k \mathbf{S}_t$ . El *efecto directo* será igual al que surge en el modelo *SDM*, pero el *efecto indirecto* se torna más simple siendo igual a la suma de los elementos fuera de la diagonal principal de  $\beta_k \mathbf{S}_t$ , conteniendo solo efectos globales.

Ya sea que se estime un modelo *SLM* o *SDM*, siempre se tendrá un efecto marginal que no es único para todas las localidades (en contraposición a lo que sucede bajo un modelo no-espacial). Debido a esta característica, LeSage y Pace (2009) proponen una medida resumen para cada uno de los efectos

considerando valores promedio. Por ejemplo para un modelo *SLM*, para cada periodo temporal se estiman los siguientes efectos:

$$\begin{aligned}\bar{M}_{Total} &= \frac{\beta_k}{n} \mathbf{1}'_n \mathbf{S} \mathbf{1}_n = \frac{\beta_k}{n} \sum_i \sum_j S_{i,j}. \\ \bar{M}_{Directo} &= \frac{\beta_k}{n} \text{tr}(\mathbf{S}) = \frac{\beta_k}{n} \sum_i S_{i,i}. \\ \bar{M}_{Indirecto} &= \frac{\beta_k}{n} \mathbf{1}'_n \mathbf{S} \mathbf{1}_n - \frac{\beta_k}{n} \text{tr}(\mathbf{S}) = \frac{\beta_k}{n} \sum_i \sum_{j \neq i} S_{i,j}.\end{aligned}$$

Estos efectos indirectos estimados sirven para conocer si los estos efectos espaciales son de importancia estadística o no. Para ello, debería estimarse los intervalos de confianza de estos efectos. El problema es que los mismos surgen de una combinación de coeficientes y puede suceder que algunos coeficientes sean significativos y otros no, sin poder determinar la significancia conjunta. LeSage y Pace (2009) proponen simular la distribución de los efectos directos e indirectos usando la matriz de varianza-covarianza implicada en la estimación por máxima verosimilitud. Si hemos estimado un *SLM*, el procedimiento consiste en extraer una combinación  $r$  de parámetros de esta matriz de varianza-covarianza:

$$\left[ \alpha_r \quad \beta'_r \quad \rho_r \quad \sigma_r^2 \right]' = \left[ \hat{\alpha} \quad \hat{\beta}' \quad \hat{\rho} \quad \hat{\sigma}^2 \right]' + P' \varepsilon,$$

donde  $P$  simboliza matriz triangular superior de la descomposición de Cholesky de la matriz de varianza-covarianza de los parámetros de máxima verosimilitud, y  $\varepsilon$  es un vector de números aleatorios de dimensión  $(K + 3)$ , provenientes de una distribución normal estándar. Si  $R$  extracciones parámetros son realizadas de esta forma, entonces los efectos promedios totales, directos e indirectos para cada variable explicativa pueden ser aproximados como:

$$\bar{\delta}_k = \frac{1}{R} \sum_{r=1}^R \delta_{kd},$$

$$t - valor_k = \bar{\delta}_k / \left[ \frac{1}{R} \sum_{r=1}^R (\delta_{kd} - \bar{\delta}_k)^2 \right],$$

donde  $\bar{\delta}_k$  es el efecto promedio de obtenido en  $R$  simulaciones, generalmente  $R = 1000$ , y dado el  $t - valor$  de los efectos, puede obtenerse la significancia de cada efecto.

Los efectos directos e indirectos desarrollados para corte transversal o paneles estáticos pueden extenderse hacia paneles dinámicos. La ventaja del modelo dinámico es que pueden obtenerse efectos directos e indirectos de corto y largo plazo. Por ejemplo, bajo un modelo *dynSDM3* (ecuación 20), los efectos de corto plazo se obtienen ignorando  $\tau$  y  $\gamma$ . La matriz de derivadas parciales de  $\mathbf{y}$  con respecto a un cambio en una unidad de la variable explicativa  $k - \acute{e}sima$  en las  $n$  regiones, para un punto temporal dado puede obtenerse como:

$$\left[ \frac{\partial E(\mathbf{y})}{\partial x_{1k}} \quad \dots \quad \frac{\partial E(\mathbf{y})}{\partial x_{nk}} \right]_t = (\mathbf{I}_n - \rho \mathbf{W})^{-1} [\beta_k \mathbf{I}_n + \theta_k \mathbf{W}]. \quad (25)$$

Para obtener los efectos de largo plazo se puede suponer se alcanza un estado estacionario ( $\mathbf{y}_t = \mathbf{y}_{t-1} = \mathbf{y}^*$ ) tal que:

$$\left[ \frac{\partial E(\mathbf{y})}{\partial x_{1k}} \quad \dots \quad \frac{\partial E(\mathbf{y})}{\partial x_{nk}} \right]_t = [(1 - \tau) \mathbf{I}_n - (\rho + \gamma) \mathbf{W}]^{-1} [\beta_k \mathbf{I}_n + \theta_k \mathbf{W}]. \quad (26)$$

Estos resultados permiten determinar los efectos directos en el corto y largo plazo, y los efectos indirectos del corto y largo plazo (*spatial spillovers*). Desde estos modelos pueden obtenerse trayectorias secuenciales periodo a periodo como las presentadas por Debarsy et al. (2012).

Es importante recordar que estos efectos deben ser cuidadosamente interpretados y no están libres de controversias ya que su obtención tiene un fuerte componente causal. La relación sobre la adecuada interpretación de los

resultados y qué se entiende por causalidad en el espacio no es un tema banal. Por ejemplo, algunos autores sugieren que el parámetro espacial  $\rho$  del *SLM* (o del *SDM*) es un parámetro causal (Gibbons y Overman, 2012). La posición más adecuada no es posiblemente esa, más bien el parámetro  $\rho$  únicamente captura la dinámica espacial de la propia variable endógena y no debería interpretarse como causal. Esto proviene de las similitudes espaciales con lo que sucede en un proceso AR(1) en el tiempo,  $y_t = \rho y_{t-1} + \varepsilon_t$ , en donde  $\rho$  nunca se interpreta como un parámetro causal en series de tiempo. Para una discusión sobre este punto véase Herrera et al. (2014).

## 5. Aplicaciones empíricas

### 5.1. Modelo de crecimiento (Ertur y Koch, 2007)

Ertur y Koch (2007) investigan la interdependencia tecnológica en el crecimiento económico entre países. Basados en un modelo neoclásico, los autores consideran una función de producción Cobb-Douglas que refleja rendimientos constantes a escala respecto al trabajo y capital físico para el país  $i$  en el periodo temporal  $t$ :

$$Y_i(t) = A_i(t)K_i^\alpha(t)L_i^{1-\alpha}(t), \quad (27)$$

donde  $Y_i(t)$  es el producto,  $K_i(t)$  es el nivel de capital físico,  $L_i(t)$  es el nivel de trabajo, y  $A_i(t)$  es el nivel agregado de tecnología especificado como:

$$A_i(t) = \Omega(t)k_i^\phi(t) \prod_{j \neq i}^n A_i^{\delta \omega_{ij}}(t). \quad (28)$$

El nivel agregado de tecnología  $A_i(t)$  de un país  $i$  depende de tres elementos. Primero, una cierta proporción de progreso tecnológico exógeno e idéntico en todos los países,  $\Omega(t) = \Omega(0)e^{\mu t}$ , donde  $\mu$  es la tasa constante de crecimiento. Segundo, el nivel tecnológico de cada país crece según el nivel agregado de capital físico por trabajador  $k_i^\phi(t) = (K_i(t)/L_i(t))^\phi$  con  $\phi \in [0; 1]$  capturando la fortaleza de las externalidades domésticas de la acumulación del capital físico. Finalmente, el tercer término captura el efecto externo de conocimiento incorporado en el capital localizado en diferentes países cuyo impacto trasciende los límites nacionales con una intensidad decreciente,  $\delta \in [0; 1]$ . Los términos  $\omega_{ij}$  representan la conectividad entre el país  $i$  y sus vecinos  $j$ .

Siguiendo el desarrollo de Solow, los autores suponen que en cada país  $i$ , una fracción constante del producto  $s_i$  es ahorrada y que la mano de obra crece exógenamente a la tasa  $l_i$ . Además, se supone una tasa anual de depreciación del capital físico constante e idéntica para todos los países. La evolución del producto por trabajador en el país  $i$  es gobernado por la conocida ecuación dinámica de Solow que permite obtener el ingreso por trabajador en el estado estacionario de cada país (Ertur y Koch, 2007, p. 1038, ecuación 9). Luego de algunas manipulaciones se obtiene el ingreso real por trabajador del país  $i$  en el estado estacionario:

$$\begin{aligned} \ln y_i^*(t) &= \frac{1}{1 - \alpha - \phi} \ln \Omega(t) + \frac{\alpha + \phi}{1 - \alpha - \phi} \ln s_i - \frac{\alpha + \phi}{1 - \alpha - \phi} \ln (l_i + g + \delta) \\ &\quad - \frac{\alpha \gamma}{1 - \alpha - \phi} \sum_{j \neq i}^n w_{ij} \ln s_j + \frac{\alpha \gamma}{1 - \alpha - \phi} \sum_{j \neq i}^n w_{ij} \ln (l_j + g + \delta) \\ &\quad + \frac{\gamma(1 - \alpha)}{1 - \alpha - \phi} \sum_{j \neq i}^n w_{ij} \ln y_j^*(t). \end{aligned} \quad (29)$$

Este es un modelo de Solow espacialmente aumentado y coincide con el predictor obtenido por Solow más un efecto espacial. Considerando  $t = 0$  por simplicidad, la anterior ecuación permite encontrar el ingreso real por

trabajador a lo largo de la trayectoria de crecimiento balanceado:

$$\ln \left[ \frac{Y_i}{L_i} \right] = \beta_0 + \beta_1 \ln s_i + \beta_2 \ln (l_i + g + \delta) + \theta_1 \sum_{j \neq i}^n w_{ij} \ln s_j + \rho \sum_{j \neq i}^n w_{ij} \ln \left[ \frac{Y_j}{L_j} \right] + \varepsilon_i, \quad (30)$$

donde  $1/(1-\alpha-\phi) \ln \Omega(0) = \beta_0 + \varepsilon_i$ ,  $i = 1, \dots, n$ , siendo  $\beta_0$  una constante y  $\varepsilon_i$  un término de error con  $\ln \Omega(0)$  reflejando la tecnología, dotación de recursos, etc. específicas de cada país. Se supone además que  $g + \delta = 0,05$ . Finalmente, se tienen las siguientes restricciones teóricas entre coeficientes:  $\beta_1 = -\beta_2 = \frac{\alpha+\phi}{1-\alpha-\phi}$  y  $\theta_2 = -\theta_1 = \frac{\alpha\gamma}{1-\alpha-\phi}$ .

La ecuación (30) coincide con el modelo *SDM* y es estimada usando información del ingreso real, inversión y crecimiento poblacional de una muestra de 91 países durante 1960 – 1995. Respecto a la matriz de pesos espaciales, se consideran dos matrices con criterios de distancia geográfica: la primera utiliza la función inversa de la distancia al cuadrado (principal matriz que refleja una aproximación gravitacional) y la segunda se basa en la función exponencial negativa de la distancia al cuadrado (utilizada como prueba de robustez en la especificación).

Ambas matrices,  $\mathbf{W}_{inv}$  y  $\mathbf{W}_{exp}$  son estandarizadas por filas tal que cada elemento es transformado como:  $\omega_{hij} = \omega_{hij}^* / \sum_{j=1}^n \omega_{hij}^*$ ;  $h = inv, exp$ , en donde

$$\omega_{inv,ij}^* = \begin{cases} 0 & \text{si } i = j \\ d_{ij}^{-2} & \text{en otro caso} \end{cases}, \quad \omega_{exp,ij}^* = \begin{cases} 0 & \text{si } i = j \\ e^{-2d_{ij}} & \text{en otro caso} \end{cases}, \quad (31)$$

siendo  $d_{ij}$  la distancia del gran-círculo entre capitales de los países  $i$  y  $j$ .

Los autores analizan varias especificaciones siguiendo restricciones teóricas sobre los parámetros, incluyendo modelos espaciales alternativos. En el Cuadro 1 se presentan los resultados del modelo de Solow tradicional estimado por *MCO* y los resultados del modelo *SDM* estimado por *ML* bajo las dos matrices

alternativas. La parte superior del Cuadro 1 muestra las estimaciones no restringidas y la parte inferior muestra los resultados bajo las restricciones teóricas  $\beta_1 = -\beta_2$  y  $\theta_2 = -\theta_1$ . Para el modelo de Solow, los resultados del test  $I$  de Moran (Moran, 1950) rechaza fuertemente la hipótesis nula de no autocorrelación espacial. Este resultado refleja el error de especificación del modelo de Solow tradicional.

Las otras columnas del Cuadro 1 muestran el modelo de Solow aumentado. En los modelos no restringidos los coeficientes tienen el signo predicho por la teoría y el coeficiente de autocorrelación espacial,  $\rho$ , es positivo y significativo. Las restricciones teóricas  $\beta_1 = -\beta_2$  y  $\theta_2 = -\theta_1$  no son rechazadas por los test  $LR$  (2,07 y 1,88, respectivamente) por lo que los modelos pueden estimarse bajo la restricciones teóricas.

Las estimaciones de los modelos restringidos confirman el error de especificación del modelo tradicional de Solow mediante los resultados del test  $I$  de Moran. Los modelos de Solow aumentado muestran un  $\alpha$  implicado cercano a un tercio y significativo. Además, el valor estimado de  $\gamma$  es cercano a 0,5 y significativo indicando la importancia de la interdependencia tecnológica entre países y la importancia del vecindario en la determinación del ingreso real ( $\rho$  estimado significativo).

Finalmente, los criterios de información  $AIC$  y  $Entropía$  muestran preferencia por el modelo bajo  $\mathbf{W}_{inv}$  y similar evidencia es mostrada por la probabilidad posterior de la estimación bayesiana. El desarrollo teórico del modelo de crecimiento guía la especificación econométrica pero no resuelve la incertidumbre sobre (1) el criterio de conectividad que define qué países son vecinos, y (2) el supuesto de pesos espaciales que define cómo los vecinos se afectan entre si. Los autores definen un criterio de conectividad basado en la distancia y plantean cierta incertidumbre sobre la fortaleza de la interacción. Ertur y Koch usaron el criterio  $AIC$  para elegir la primera matriz y, en este ejemplo, se han agregado dos alternativas que confirman esa elección.

Cuadro 1: Estimaciones alternativas

<i>Modelos no-restringidos</i>	<i>Solow</i>	<i>SDM/W<sub>inv</sub></i>	<i>SDM/W<sub>exp</sub></i>
constante	4,65**	0,48	1,81
ln( <i>s</i> )	1,28***	0,82***	0,80***
ln( <i>l</i> + 0,05)	-2,71***	-1,50***	-1,58***
<b>W</b> × ln( <i>s</i> )		-0,28	-0,24
<b>W</b> × ln( <i>l</i> + 0,05)		0,24	0,52
<b>W</b> × ln( <i>y</i> )		0,71***	0,62***
<b>I Moran</b> ( <b>W<sub>inv</sub></b> )	0,41***		
<b>I Moran</b> ( <b>W<sub>exp</sub></b> )	0,40***		
<i>Modelos restringidos</i>			
constante	8,38***	2,32***	3,11***
ln( <i>s</i> ) - ln( <i>l</i> + 0,05)	1,38***	0,85***	0,82***
<b>W</b> × [ln( <i>s</i> ) - ln( <i>l</i> + 0,05)]		-0,22	-0,19
<b>W</b> × ln( <i>y</i> )		0,71***	0,62***
<b>I Moran</b> ( <b>W<sub>inv</sub></b> )	0,40***		
<b>I Moran</b> ( <b>W<sub>exp</sub></b> )	0,38***		
<i>Test de restricciones</i>			
<i>α</i> implicado	4,43**	2,07	1,88
<i>α</i> implicado	0,58***	0,28**	0,30**
<i>φ</i> implicado		0,18*	0,15
<i>γ</i> implicado		0,56***	0,51***
<i>AIC</i>		97,54	105,37
<i>Entropía</i>		30,83	34,74
<i>Bayes (prob. post.)</i>		0,98	0,02

**Nota:** \*, \*\* y \*\*\* denota significancia al 10%, 5% y 1% respectivamente. Código Matlab disponible en: <https://sites.google.com/site/mherreraagomez/principal/Codes>.



## 5.2. Consumo de tabaco en datos de panel

Baltagi y Li (2004) explican la demanda de cigarrillos en 46<sup>3</sup> estados norteamericanos en un periodo de 30 años, desde 1963 a 1992. La variable dependiente,  $\log(c)$ , es el logaritmo de la venta per cápita de cigarrillos en términos reales, medida en paquetes de cigarrillos por personas de 14 o más años. Las variables explicativas son:  $\log(p)$ , logaritmo del precio de venta promedio del paquete de tabaco en el estado y  $\log(ing)$ , logaritmo del ingreso disponible per cápita del estado. La matriz de pesos espaciales es especificada como una matriz de contigüidad binaria, estandarizada por fila. Esta base de datos ha sido utilizada por Debarsy et al. (2012), Elhorst (2014a) y Kelejian y Piras (2014), entre otros.

La motivación inicial del trabajo fue el efecto contrabando, en donde compradores de cigarrillos de estados limítrofes compraban cigarrillos en los estados vecinos si el precio era ventajoso. Esto podría ser para consumo propio o para revender en su estado de origen.

El Cuadro 2 presenta los resultados de modelos estáticos y dinámicos. De los dos modelos estáticos, el modelo *pooled* permite una primera aproximación para detectar la presencia de elementos faltantes como efectos espaciales, sustantivos y residuales. Esta información es obtenida por las versiones robustas de los tests  $LM$ ,  $LM'_{error}$  y  $LM'_{lag}$ , que detectan la omisión de un efecto espacial en el error o un efecto espacial en la variable dependiente, respectivamente. Ambos tests rechazan la hipótesis nula de no autocorrelación residual y sustantiva, permitiendo avanzar en la especificación espacial. Adicionalmente, se testea en el modelo *pooled* la significancia conjunto de efectos fijos individuales y temporales encontrando evidencia a favor de la inclusión de ambos efectos.

---

<sup>3</sup>Los estados omitidos son Alaska, Colorado, Hawaii, Carolina del Norte y Oregon.

Considerando los resultados del modelo *pooled*, la especificación del modelo puede seguir la secuencia de abajo hacia arriba de la Figura 1. En nuestro caso aplicaremos la estrategia de utilizar el modelo estático más complejo posible, *SDM*, y a partir de allí reducir elementos no significativos.

Los resultados del modelo estático *SDM* se presentan en la tercera columna del Cuadro 2. La hipótesis nula  $\theta = 0$  permite testear si el *SDM* puede ser simplificado a un *SLM*, el test de *Wald* rechazar dicha hipótesis. La posible reducción del *SDM* a un *SEM* ( $H_0 : \theta + \rho\beta = 0$ ) es rechazada por el test de factores comunes. Finalmente, la significancia del parámetro  $\hat{\rho}$  muestra que no es posible reducir el modelo *SDM* en un *SLX*.

Cuadro 2: Estimaciones alternativas

Modelos	Estáticos		Dinámicos	
	<i>Pooled</i>	<i>SDM</i>	<i>dynSDM1</i>	<i>dynSDM4</i>
$\log(c)_{-1}$			0,87***	0,87***
$\mathbf{W} \times \log(c)_{-1}$				0,02
$\mathbf{W} \times \log(c)$		0,23***	0,05**	0,03
$\log(p)$	-0,86***	-1,00***	-0,26***	-0,26***
$\log(y)$	0,27***	0,60***	0,10***	0,10***
$\mathbf{W} \times \log(p)$		0,06	0,16***	0,16***
$\mathbf{W} \times \log(y)$		-0,29***	-0,02	-0,02
$LM'_{error}$	144,84***			
$LM'_{lag}$	58,26***			
$Wald(\theta = 0)$		16,06***		
$Wald(fact. com.)$		8,62**		

**Nota:** \*, \*\* y \*\*\* denota significancia al 10%, 5% y 1% respectivamente. Modelos *SDM* incluyen FE.

Una vez detectado el modelo estático final, puede procederse a interpretar los resultados. Los coeficientes de las dos variables explicativas tienen el signo esperado con elasticidad unitaria negativa para el precio y de 0.6 positiva para el ingreso indicando que un mayor ingreso aumenta la demanda de cigarrillos. Comparados estos coeficientes con los obtenidos por el modelo *pooled* parece claro que existe una corrección, sin embargo, esta comparación no es válida debido a la presencia del parámetro  $\rho$ . Es necesario computar los efectos marginales para proveer de una interpretación adecuada. En el Cuadro 3 se muestran los efectos marginales del modelo estático *SDM* basados en 1000 repeticiones. Los resultados son interpretados como de largo plazo, una vez que la dinámica espacial del consumo converge. El efecto de principal interés es el impacto marginal indirecto del precio y como puede observarse el resultado es una elasticidad-precio  $-0,21$  y significativo al 5%, evidencia que no es consistente con la hipótesis inicial de Baltagi y Lee. Esto puede ser explicado por la ausencia de elementos dinámicos temporales.

Para conocer si existe dinámica temporal, se han estimado dos versiones dinámicas del modelo *SDM*: *dynSDM1* (ecuación 18) y *dynSDM4* (ecuación 21), ambos considerando efectos fijos individuales y temporales. Como se muestra en el Cuadro 3, ambos modelos detectan una alta persistencia en el consumo de cigarrillos con un coeficiente del  $\ln(c)_{-1}$  de 0.87. Sin embargo, el modelo *dynSDM4* no detecta persistencia espacio-recursiva siendo  $\mathbf{W} \times \ln(c)_{-1}$  no significativo. Por lo tanto, el modelo más adecuado es el *dynSDM1*.

Bajo este modelo final, *dynSDM1*, nuevamente se obtienen los efectos marginales de las variables explicativas que se presentan en el Cuadro 3. Debido a la presencia de dinámica temporal es posible obtener efectos marginales de corto y largo plazo. Centrando el interés en el efecto indirecto del precio, puede deducirse que el efecto contrabando en el corto plazo, elasticidad-precio 0,15 y significativa: el incremento del precio en estados vecinos aumenta el consumo en cada región. Este efecto en el largo plazo continúa siendo positivo pero no es

significativo. En conclusión, existe evidencia de *spillovers* espaciales del precio del cigarrillo a corto plazo.

Cuadro 3: Efectos directos e indirectos de modelos *SDM* con efectos fijos

<i>Modelos</i>	<i>SDM</i>	<i>dynSDM1</i>
<i>Corto plazo</i>		
efecto directo $\log(p)$		-0,26***
efecto directo $\log(y)$		0,10***
efecto indirecto $\log(p)$		0,15***
efecto indirecto $\log(y)$		-0,02
<i>Largo plazo</i>		
efecto directo $\log(p)$	-1,01***	-1,94***
efecto directo $\log(y)$	0,59***	0,78***
efecto indirecto $\log(p)$	-0,21**	0,59
efecto indirecto $\log(y)$	-0,19**	0,25

**Nota:** \*, \*\* y \*\*\* denota significancia al 10%, 5% y 1% respectivamente. Código Stata disponible en: <https://sites.google.com/site/mherreraomez/principal/Codes>.

## 6. Comentarios finales

El desarrollo de la econometría espacial ha sido muy importante en los últimos años acompañando la mayor disponibilidad de información georreferenciada. El capítulo presenta los desarrollos más consolidados dentro del área para datos de corte transversal y datos de panel destacando tópicos relativos a

la especificación e interpretación de los modelos espaciales con dependencia espacial global (no se aborda el tema de la heterogeneidad espacial que da lugar a tratamientos locales).

Los datos espaciales presentan características particulares. Los datos pueden ser continuos o discretos, pueden provenir de agregaciones u observaciones puntuales en el espacio, la localización puede ser regular o irregular, y dichas localizaciones pueden ser consideradas provenientes de un espacio continuo o discreto. La agregación altera la variabilidad e impacta en la inferencia a realizar. También el soporte de la información geográfica es importante cuando se combinan diferentes tipos de datos espaciales. Estas características deben ser consideradas y las conclusiones inferenciales deberían realizarse con suma cautela teniendo en cuenta la escala de análisis, más aún si estas conclusiones son orientadas a acciones políticas.

Un tema central en esta disciplina es la construcción de la matriz de pesos espaciales y los problemas inferenciales que puede generar una mala elección. La normalidad asintótica y consistencia de los estimadores está bien establecida bajo el supuesto de una matriz espacial estrictamente exógena. Sin embargo, cuando los pesos espaciales son endógenos los métodos no son válidos. Algunos recientes desarrollos flexibilizan este supuesto, por ejemplo, Kelejian y Piras (2014) proponen una matriz de pesos espaciales endógenos instrumentalizada. Cheng y Lee (2017) utilizan esta estrategia para desarrollar un test de endogeneidad tipo Hausman para los pesos espaciales. Qu et al. (2017) y Qu y Lee (2017) desarrollan enfoques de estimación permitiendo correlación entre el término error y las variables utilizadas en la construcción de los pesos espaciales. Esta endogeneidad es solucionada utilizando el método de función de control. Estos modelos son actualmente de frontera y existen varios avances que flexibilizan la matriz de contactos permitiendo que la misma cambie entre periodos temporales.

Los ejemplos empíricos presentados en el capítulo pretenden mostrar dos vías alternativas habituales en econometría espacial aplicada. El trabajo de Ertur y Koch (2007) es un buen ejemplo de la aplicación del enfoque *theory-driven* y cómo la estimación final íntegramente vincula la interpretación a la teoría de crecimiento económico. A pesar de ser un modelo derivado desde la teoría, la especificación de la matriz es un a priori de los investigadores y por ello la necesidad de comparar matrices alternativas. Esta incertidumbre está presente en todo modelo espacial y existen diferentes alternativas que ayudan a la elección de la matriz más adecuada (véase Herrera et al., 2017). El segundo caso aplicado es un ejemplo del enfoque *data-driven*, explorando especificaciones espaciales alternativas y con conclusiones obtenidas del modelo final. En este caso no hay un desarrollo teórico que fundamente el modelo finalmente estimado, son los datos los que permiten adecuar la mejor especificación. La interpretación de los resultados son realizados según la vigente visión de efectos marginales simulados. La interpretación de estos efectos marginales posee un fuerte significado causal que no deja de ser controvertido y es un tema que merece un mayor debate a futuro.

Finalmente, cabe mencionar las diferentes alternativas computacionales. Bajo Matlab existe una amplia cantidad de códigos que cubren todo el rango de modelos presentados en el capítulo. El programa *R* tiene diferentes paquetes que permiten estimar modelos de corte transversal y paneles espaciales estáticos. GeoDa es un excelente programa para iniciarse en el análisis espacial, permitiendo la estimación de modelos de corte transversal y paneles *pooled* bajo *ML*. GeoDaSpace es útil si se desea estimar modelos de corte transversal por *IV/GMM*. También existe un conjunto de opciones desarrolladas bajo Python. Por último, la versión 15 de Stata trae incorporado un módulo para análisis espacial y hay una nutrida variedad de comandos desarrollados por usuarios que cubren todo el espectro de los modelos espaciales aquí analizados.

## Referencias

- Anselin, L. (1988). *Spatial econometrics: Methods and models*, Volume 4. Kluwer Academic Publishers, Dordrecht.
- Anselin, L. (2005). Exploring spatial data with GeoDa: A workbook. Urbana-Champaign: Center for Spatially Integrated Social Science.
- Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1), 3-25.
- Anselin, L., A. Bera, R. Florax, y M. Yoon (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26(1), 77-104.
- Anselin, L., J. Le Gallo, y H. Jayet (2008). Spatial panel econometrics. En L. Matyas y P. Sevestre (Eds.), *The econometrics of panel data: fundamentals and recent developments in theory and practice*, vol. 46, 625-660. Springer.
- Arbia, G. (2012). *Spatial data configuration in statistical analysis of regional economic and related problems* (Vol. 14). Springer Science & Business Media.
- Arraiz, I., D. Drukker, H. Kelejian, e I. Prucha (2010). A spatial Cliff-Ord-type model with heteroskedastic innovations: Small and large sample results. *Journal of Regional Science*, 50(2), 592-614.
- Baltagi, B. y D. Li (2004). Prediction in the panel data model with spatial correlation. En L. Anselin, R. Florax, y S. Rey (Eds.), *Advances in spatial econometrics*, 283-295. Springer.
- Baltagi, B., S. Song, y W. Koh (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics*, 117(1), 123-150.
- Bartels, C. y R. Ketellapper (1979). *Exploratory and explanatory statistical analysis of spatial data*. Martinus Nijho, Boston, MA.
- Bennett, R. (1979). *Spatial time series*. Pion, London.
- Bivand, R. (2010). Exploratory spatial data analysis. En M. Fischer y G (Eds.), *Handbook of Applied Spatial Analysis*, 219-254. Springer, Hiedelberg.

- Brock, W. y S. Durlauf (1995). Discrete choices with social interactions I: Theory. Santa Fe, Santa Fe Institute WP 95-10.
- Brueckner, J. (2003). Strategic interaction among governments: An overview of empirical studies. *International Regional Science Review*, 26(2), 175-188.
- Cheng, W. y L.-f. Lee (2017). Testing endogeneity of spatial and social networks. *Regional Science and Urban Economics*, 64, 81-97.
- Cliff, A. y J. Ord (1981). *Spatial Processes: Models and Applications*. Pion: London.
- Cressie, N. (1993). *Statistics for spatial data*. New York, Wiley.
- Debarsy, N. y C. Ertur (2010). Testing for spatial autocorrelation in a fixed effects panel data model. *Regional Science and Urban Economics*, 40(6), 453-470.
- Debarsy, N., C. Ertur, y J. P. LeSage (2012). Interpreting dynamic space-time panel data models. *Statistical Methodology*, 9(1), 158-171.
- Drukker, D. , P. Egger, e I. Prucha (2013). On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. *Econometric Reviews*, 32(5-6), 686-733.
- Elhorst, J. (2010a). Dynamic panels with endogenous interaction effects when T is small. *Regional Science and Urban Economics*, 40(5), 272-282.
- Elhorst, J. (2010b). Spatial panel data models. En M. Fischer y A. Getis (Eds.), *Handbook of Applied Spatial Analysis. Software Tools, Methods and Applications*, 377-407. Springer-Verlag Berlin Heidelberg.
- Elhorst, J. (2014a). Matlab software for spatial panels. *International Regional Science Review*, 37(3), 389-405.
- Elhorst, J. (2014b). *Spatial Econometrics. From Cross-sectional data to Spatial Panels*. SpringerBriefs in Regional Science, Springer.
- Ertur, C. y W. Koch (2007). Growth, technological interdependence and spatial externalities: Theory and evidence. *Journal of Applied Econometrics*, 22(6), 1033-1062.



- Giacomini, R. y C. Granger (2004). Aggregation of space-time processes. *Journal of Econometrics*, 118(1), 7-26.
- Gibbons, S. y H. Overman (2012). Mostly pointless spatial econometrics?. *Journal of Regional Science*, 52(2), 172-191.
- Gotway, C. y L. Young (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458), 632-648.
- Graubard, B. y E. Korn (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17(1), 73-96.
- Griffith, D. A. (1996). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. En S. Arlinghaus (Ed.), *Practical Handbook of Spatial Statistics*, p. 65-82. CRC Press Book: Boca Raton, FL.
- Halleck Vega, S. y J. Elhorst (2015). The SLX model. *Journal of Regional Science*, 55(3), 339-363.
- Hendry, D. (1979). Predictive failure and econometric modelling in macroeconomics: The transactions demand for money. En P. Ormerod (Ed.), *Economic Modelling: Current Issues and Problems in Macroeconomic Modelling in the UK and the US*, 217-242. Heinemann Education Books, London.
- Herrera, M. (2015). Econometría espacial usando Stata. Breve guía aplicada para datos de corte transversal. Documentos de Trabajo del IELDE (13).
- Herrera, M., J. Mur, y M. Ruiz (2016). Detecting causal relationships between spatial processes. *Papers in Regional Science*, 95(3), 577-594.
- Herrera, M., J. Mur, y M. Ruiz (2017). A comparison study on criteria to select the most adequate weighting matrix. XIth World Conference of the Spatial Econometrics Association, Singapore.
- Hordijk, L. (1979). Problems in estimating econometric relations in space. *Papers in Regional Science*, 42(1), 99-115.

- Kapoor, M., H. Kelejian, e I. Prucha (2007). Panel data models with spatially correlated error components. *Journal of Econometrics*, 140, 97-130.
- Kelejian, H. y D. Robinson (1995). Spatial correlation: a suggested alternative to the autoregressive model. En L. Anselin y R. Florax (Eds.), *New directions in Spatial Econometrics*, 75-95. Springer.
- Kelejian, H. y G. Piras (2014). Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. *Regional Science and Urban Economics*, 46, 140-149.
- Kelejian, H. e I. Prucha (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1), 99-121.
- Kelejian, H. e I. Prucha (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40(2), 509-533.
- Kelejian, H. e I. Prucha (2010). Specication and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1), 53-67.
- Krugman, P. (1991). Increasing returns and Economic Geography. *Journal of Political Economy*, 99(3), 483-499.
- Lee, L.-f. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, 22(4), 307-335.
- Lee, L.-f. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6), 1899-1925.
- Lee, L.-f. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140(2), 333-374.

- Lee, L.-f. y J. Yu (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, 154(2), 165-185.
- Lee, L.-f. y J. Yu (2014). Spatial panel models. En B. Baltagi (Ed.), *The Oxford Handbook of Panel Data*, 363-401. Oxford University Press.
- LeSage, J. y R. Pace (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC press.
- Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3), 531-542.
- Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17-23.
- Openshaw, S. (1984). *The modifiable areal unit problem*. Geo Books, Norwich. Printed by Headley Brothers Ltd. Kent.
- Openshaw, S. y P. Taylor (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. En N. Wrigley y P. Taylor (Eds.), *Statistical Methods in the Spatial Sciences*, 127-144. Pion.
- Ord, J. (1981). Towards a theory of spatial statistics: a comment. *Geographical Analysis*, 13(1), 86-91.
- Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349), 120-126.
- Pace, R. y S. Zhu (2012). Separable spatial modeling of spillovers and disturbances. *Journal of Geographical Systems*, 14(1), 75-90.
- Paelinck, J. y L. Klaassen (1979). *Spatial Econometrics*. Saxon House, Farnborough.
- Pesaran, M. (2004). General diagnostic tests for cross section dependence in panels. Cambridge Working Papers in Economics No. 0435. University of Cambridge.
- Qu, X. y L.-F. Lee (2017). Estimation of an endogenous SAR model with spatial weights constructed by bilateral variables. XIth World Conference of the Spatial Econometrics Association, Singapore.

- Qu, X., L.-f. Lee, y J. Yu (2017). QML estimation of spatial dynamic panel data models with endogenous time varying spatial weights matrices. *Journal of Econometrics*, 197(2), 173-201.
- Smith, T. (2009). Estimation bias in spatial models with strongly connected weight matrices. *Geographical Analysis*, 41(3), 307-332.
- Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46, 234-240.
- Waller, L. y C. Gotway (2004). *Applied Spatial Statistics for Public Health Data*, v. 368. John Wiley & Sons.
- Yu, J., R. de Jong, y L.-f. Lee (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *Journal of Econometrics*, 146 (1), 118-134.
- Yu, J., R. de Jong, y L.-f. Lee (2012). Estimation for spatial dynamic panel data with fixed effects: The case of spatial cointegration. *Journal of Econometrics*, 167(1), 16-37.

## **Capítulo 5. Big data y aprendizaje automático: Ideas y desafíos para economistas**

Walter Sosa Escudero

### **1. Introducción**

Basta ya de la teoría del comportamiento humano, desde la lingüística a la sociología. Olvídense de la taxonomía, la ontología y la psicología. ¿Quién sabe por qué la gente hace lo que hace? El punto es que lo hacen, y podemos monitorearlo y medirlo con una fidelidad sin precedentes. Con suficientes datos los números hablan por sí mismos (Chris Anderson, *Wired*, 23/5/2008).

Con llamativa frecuencia aparecen ideas revolucionarias destinadas a barrer con el canon de una disciplina. La epistemología ha dedicado un espacio considerable al estudio de esta dinámica discontinua y al de las resistencias internas de las ciencias a abandonar sus paradigmas.

La econometría no estuvo exenta a esta evolución no lineal. Así, la econometría dinámica de la escuela de Hendry, el enfoque de cointegración, el método generalizado de momentos, los modelos estructurales en organización industrial empírica o los métodos experimentales y cuasi experimentales llegaron acompañados de una sensación de cambio permanente, siempre con un planteo (explícito o no) de revolución.

Las ideas asociadas al título de este trabajo (*big data*, aprendizaje automático, minería de datos, etc.) traen una propuesta concreta de revisión sustancial de los paradigmas metodológicos de la ciencia, a la luz del virtual “diluvio de datos” provocado por las redes sociales y el uso de dispositivos interconectados, y que en sus visiones más extremas prometen cambios radicales como los pronosticados por Chris Anderson en la cita que da comienzo a este trabajo, o, en el otro extremo la mera sensación de que se trata de otra moda pasajera.

Si bien lejos del escepticismo extremo, llama la atención que las ideas de *big data* o aprendizaje automático encuentren cierta resistencia en la economía y las ciencias sociales en general, sobre todo en comparación con la actitud optimista prevaleciente en casi todo el espectro científico.

Este trabajo revisa algunas ideas básicas del paradigma de *big data* y aprendizaje automático con el doble propósito de diseminar el uso de estas tecnologías, discutir sus ventajas y desventajas y reflexionar sobre las razones que explican la cautela imperante en economía. Se trata de una introducción técnicamente simple, que puede ser pensada como un capítulo adicional al de un texto básico de econometría como el de Wooldridge (2015) o el de Stock y Watson (2016). El capítulo presupone conocimientos del paradigma clásico en econometría, en particular del modelo lineal estimado por el método de mínimos cuadrados. El capítulo cubre una *mínima* cantidad de material, suficiente para apreciar las principales diferencias entre el enfoque de aprendizaje automático y el clásico de la econometría, observar las principales

desventajas y ventajas y, fundamentalmente, motivar e iniciar al interesado en esta literatura, priorizando intuiciones y conceptos por sobre tecnicismos.

El capítulo se estructura de la siguiente forma. La sección 2 discute las principales diferencias entre la econometría clásica y el enfoque de *machine learning*. A esta altura conviene advertir (y tal vez pedir disculpas) sobre algún abuso de terminología en inglés, toda vez que no existen traducciones obvias de muchos neologismos usados en esta temática. La regla a utilizar es preferir términos en español siempre y cuando su uso no lleve a confusiones innecesarias. La sección 3 introduce algunas ideas básicas (construcción de modelos, complejidad, validación cruzada, etc.) dentro de un contexto familiar para los econométricos: el modelo de regresión lineal. La idea es extendida en la dirección del problema de clasificación en la sección 4, que ocupa un espacio considerable en esta literatura. La sección 5 discute algunas estrategias de *aprendizaje no supervisado*, como la construcción de clusters o el problema de reducción de la dimensionalidad. La sección 6 reflexiona sobre la reticencia a adoptar técnicas de *big data* en economía, a la luz del paradigma experimental/estructural que dominó la econometría desde sus inicios. La sección 7 ofrece una breve guía de lectura para los interesados en ahondar en esta temática.

## 2. Aprendizaje, estadística y econometría

A fines de resaltar la comparación entre enfoques, es conveniente concentrar la discusión en la herramienta por excelencia de la econometría: el modelo lineal estimado por el método de mínimos cuadrados ordinarios:

$$y_i = x_i' \beta + u_i, \quad i = 1, \dots, n,$$

en donde  $x_i$  es un vector de  $K$  variables explicativas,  $\beta$  es un vector de  $K$  coeficientes y  $u_i$  es el término de error del modelo, que representa factores no observables que impactan en  $y_i$  pero que no son captados por  $x_i$ , para una muestra de  $i = 1, \dots, n$  observaciones. El estimador mínimo cuadrático de  $\beta$  es  $\hat{\beta} = (\sum_{i=1}^n x_i x_i')^{-1} (\sum_{i=1}^n x_i y_i)$ .

El modelo lineal es una herramienta clave en economía y varias ciencias sociales, “el automóvil de la estadística moderna” a decir de Stephen Sigler, el prestigioso historiador de la estadística. Explica su enorme popularidad el hecho de que bajo algunos supuestos  $\beta$  contiene información causal acerca de como  $x$  afecta a  $y$  (en particular, a su valor esperado). Tal vez la principal contribución del análisis de regresión lineal es que permite al científico social medir efectos causales sin necesidad de recurrir a experimentos explícitos. Mas concretamente, el modelo lineal reemplaza experimentación por control: mientras que el experimento mueve causas de a una por vez, el modelo de regresión *controla* por la presencia de factores que pueden confundir el efecto puro de una causa con de otros factores concurrentes, sin requerir que estos se hayan movido de a uno.

No es exagerado decir que la historia de la econometría puede escribirse a la luz de los intentos de justificar el modelo de regresión (y sus variantes) como una herramienta para la estimación confiable de efectos causales, ya sea como reemplazo de un experimento explícito o como representación de una teoría económica, en una línea argumental que va desde los problemas de identificación en modelos de ecuaciones simultáneas a los métodos cuasi experimentales de la así llamada “revolución de credibilidad” de las últimas dos décadas, pasando por el método de variables instrumentales o el problema de selectividad muestral, por mencionar algunas contribuciones de la econometría al análisis empírico.

Lo cierto es que en esta saga el actor fundamental es el vector de coeficientes desconocido  $\beta$ , que cuantifica la forma en la que  $x$  afecta a  $y$ , de modo que una



parte sustancial de la econometría se dedica a establecer condiciones bajo las cuales es posible recuperar  $\beta$  bajo una variedad de condiciones poblacionales y muestrales.

## 2.1. El paradigma inferencial frecuentista en econometría

En la visión frecuentista (en oposición a bayesiana) dominante en econometría el principal objetivo consiste en estimar  $\beta$  de la mejor manera posible, a la vez de garantizar mecanismos que permitan realiza “inferencias válidas” entendidas como la construcción de intervalos de confianza, la medición de la imprecisión implícita en la estimación o la implementación de tests de hipótesis.

Focalizando en el problema de estimación, en términos muy generales el problema de la evaluación de la calidad de un estimador  $\hat{\beta}$  se reduce a cotejar cuan “lejos” está de su objetivo  $\beta$ . Dicha comparación es compleja toda vez que  $\hat{\beta}$  es una variable aleatoria y  $\beta$  no. Así, la noción de insesgadez reduce el problema a comparar  $E(\hat{\beta})$  con  $\beta$ , dos valores determinísticos.

El famoso Teorema de Gauss Markov establece que bajo ciertas condiciones sobre la estructura probabilística del modelo lineal general, es posible aseverar que  $\hat{\beta}$  es óptimo si bien en un sentido muy limitado:  $\hat{\beta}$  tiene varianza mínima en la familia de estimadores lineales e insesgados.

Es crucial remarcar que asociar la calidad de un estimador a la verificación de ciertas propiedades (insesgadez, varianza mínima, etc.) es una tarea *ex-ante*, es decir, depende de que valgan ciertos supuestos que no necesariamente pueden evaluarse empíricamente. Mas concretamente, una propiedad como la de insesgadez es una propiedad del *estimador* (entendido como una regla

o algoritmo) y no de la *estimación* (lo que surge de cargar los datos en el algoritmo).

Consecuentemente, la evaluación de la calidad de un procedimiento estadístico (un estimador, un test de hipótesis) requiere postular un modelo (un proceso generador de datos) contra el cual cotejarlo. La práctica generalizada de la econometría frecuentista toma al modelo como dado y focaliza su energía en como estimarlo de la mejor manera posible. Es decir, el modelo no se construye dentro del proceso de estimación o inferencia sino que viene dado exógenamente. Usualmente la relación  $y_i = x_i' \beta + u_i$  proviene de alguna teoría o “estructura” que relaciona a  $y$  con  $x$  y  $u$ , o de la descripción de un experimento en donde las variaciones en  $x$  se traducen en  $y$ .

La econometría clásica tiende a limitarse a proveer estrategias de estimación e inferencia óptimas de un modelo dado y en algún contexto relevante. A modo de ejemplo, si  $x_i$  es aleatoria y  $E(u_i|x_i) = 0$ ,  $\hat{\beta}$  es en general un estimador consistente. O, adicionalmente, el uso del estimador de Eicker-White para la varianza de  $\hat{\beta}$  permite realizar inferencia válida independientemente de que valga o no el supuesto de homocedasticidad, siempre y cuando el tamaño de la muestra sea lo suficientemente grande.

Cabe preguntarse cuál es el rol de la cantidad de datos en el enfoque frecuentista paradigmático de la econometría, y en ese sentido, cual es la contribución de *big data* en la dirección de “muchos datos”. El grueso de la práctica favorece, casi lexicográficamente, estimadores insesgados o consistentes. Por razones complejas de entender cualquier otra propiedad de un estimador (eficiencia, robustez, etc.) esta subsumida a que el estimador este centrado en el verdadero valor de los parámetros. Entonces, si los datos provienen del mismo modelo asumido y son lo suficientemente independientes y homogéneos, mas datos se traducen en estimaciones más eficientes (o tests con mayor potencia) o dotan de mayor confiabilidad a las aproximaciones

asintóticas, cuya validez depende de que el tamaño de la muestra sea lo suficientemente grande.

A la luz de la fuerte preferencia por la insesgadez, la historia reciente de la econometría (los últimos 30 años) ha priorizado contextos empíricos en donde los supuestos que garantizan esta propiedad valen, mas allá de la cantidad de datos. A modo de ejemplo, si los datos provienen de un experimento bien diseñado,  $E(u_i|x_i) = 0$  por construcción ya que la variación experimental en  $x_i$  no interactúa con  $u_i$ . Alternativamente, ciertos episodios institucionales generan variaciones en  $x$  y  $y$  que tienen esta propiedad (los así llamados experimentos naturales), o garantizan la existencia de una variable que puede funcionar como un instrumento válido.

Entonces, la práctica econométrica ha priorizado el uso de datos en donde para el problema de interés se puede garantizar cierta variación exógena que permita aislar el efecto de  $x$  sobre  $y$ . Y en pos de este objetivo, se han preferido bases de datos relativamente pequeñas por sobre otras mas masivas pero en donde no es posible garantizar supuestos de exogeneidad. En todo caso, la profusión de datos no fue explotada para construir modelos mas complejos sino para aislar subconjuntos de datos donde existen garantías mínimas de estimación insesgada o consistente. Habiendo sorteado esta barrera inicial, la existencia de mas datos cumple un rol pasivo, a través de ganancias de eficiencia o dando mas credibilidad a las aproximaciones asintóticas.

## 2.2. Aprendizaje automático y construcción de modelos

La idea de aprendizaje automático o estadístico cambia por completo la visión descrita anteriormente, al centrar la atención en  $Y$  y no en el modelo. En este enfoque el principal objetivo es predecir  $y$  dado  $x$ , de modo que el modelo

es una suerte de actor secundario introducido solo a los efectos de mejorar la *performance* predictiva.

Cuando el interés es y parece natural que la evaluación del procedimiento estadístico usado recaiga en la comparación entre  $y$  y  $\hat{y}$  (su versión estimada o predicha), en forma análoga a la comparación anterior entre  $\beta$  y  $\hat{\beta}$ .

Una forma simple de realizar esa evaluación es a través del *error cuadrático medio* de predecir  $y = y_0$

$$ECM(y_0) = E(\hat{y}_0 - y_0)^2.$$

Un importante resultado es el siguiente:

$$ECM(y_0) = V(\hat{y}_0) + [\text{Sesgo}(\hat{y}_0)]^2,$$

en donde  $\text{Sesgo}(\hat{y}_0) = E(\hat{y}_0) - y_0$ . Naturalmente, si el proceso de predicción es insesgado el problema de minimizar el *ECM* pasa exclusivamente por minimizar la varianza. Una de las principales contribuciones de la visión de aprendizaje automático es explotar la idea de que tolerar un sesgo pequeño puede implicar una dramática caída en la varianza, y tal vez un *ECM* menor que el que se obtendría restringiendo el análisis a predictores insesgados como en la econometría clásica. Esta tolerancia (o preferencia) por procedimientos sesgados es una de las principales diferencias con el paradigma de la econometría tradicional.

Pero la principal diferencia radica en que el objetivo central del aprendizaje automático es predecir correctamente *fuera de la muestra*. Es decir, el objetivo consiste evaluar la capacidad predictiva del modelo para datos que no son los usados para estimarlo. A fines de enfatizar la comparación con la práctica habitual de la econometría, por construcción el estimador MCO maximiza el coeficiente  $R^2$ , que mide la capacidad de ajuste del modelo a los datos usados para realizar la estimación, de modo que MCO maximiza la capacidad predictiva *dentro de la muestra* para un modelo dado.

Cuando el objetivo es predecir y la esencia del aprendizaje automático es *construir el modelo* a fines de maximizar la capacidad predictiva fuera de la muestra. Es crucial notar la diferencia con respecto al enfoque clásico de la econometría. En el contexto del modelo lineal estimado por MCO, si el objetivo fuese maximizar la capacidad predictiva dentro de la muestra, cualquier modelo que regrese y en  $K = n$  variables explicativas no colineales produce un  $R^2 = 1$ , es decir, el modelo predice sin error los datos observados. El desafío de *machine learning* es predecir *fuera* de la muestra, lo cual elimina estrategias triviales (llamadas de “sobreajuste”) como la discutida anteriormente.

Entonces, la diferencia fundamental con el enfoque clásico es que aprendizaje automático o estadístico es fundamentalmente una estrategia de *construcción de modelos* (y no de estimación) con el objetivo de predecir correctamente fuera de la muestra. Consecuentemente, el primer desafío consiste en definir con alguna precisión a) que significa “construir un modelo”, b) como se evalúa la capacidad predictiva del mismo fuera de la muestra. A la luz de estas cuestiones, la esencia de *machine learning* reside en interactuar entre estas dos actividades: el modelo se construye y reconstruye a la luz de su capacidad predictiva. Esta interacción entre datos, criterios y modelos es lo que da sustento a la idea de “aprendizaje”.

La siguiente sección discute algunas ideas básicas de cómo funciona este procedimiento iterativo de construcción, evaluación y reconfiguración de modelos.

### 3. Regresión para la predicción

#### 3.1. Estimar vs. construir modelos

Un resultado clásico en probabilidad es que dada una variable aleatoria  $y$ ,  $E(y)$  es su mejor predictor en error cuadrático medio, es decir,  $E(y)$  es el valor de  $m$  que minimiza  $E(y - m)^2$ .

Otro resultado importante dice que para cualquier vector aleatorio  $x$  siempre es posible expresar a  $y$  de la siguiente forma

$$y = E(y|x) + u,$$

en donde  $u$  es una variable aleatoria que satisface  $E(u|x) = 0$ . Este resultado dice intuitivamente que  $y$  siempre puede descomponerse en dos partes: una relacionada con el mejor predictor en base a  $x$  ( $E(y|x)$ ) y otra ortogonal a esta y basada en factores no observables. Entonces el problema de predecir  $y$  en base a  $x$  pasa por estimar  $E(y|x)$ , entendida como el mejor predictor (en ECM) de  $y$ . El término  $u$  es la parte *irreducible* de  $y$  con respecto a la cual nada puede hacerse ya que se basa en factores no observables.

Es importante mencionar que si bien siempre es posible descomponer a  $y$  como la suma de dos términos no relacionados como  $E(y|x)$  y  $u$ , esto no implica que  $E(y|x)$  sea una relación causal que describe la forma en la que movimientos exógenos de  $x$  son capaces de alterar  $y$ . Es decir,  $E(y|x)$  contiene información predictiva sobre  $y$ , no necesariamente causal, para lo cual habría que explicitar que significa que una cosa cause a la otra. La idea de “causalidad por contrafactuales” es una alternativa posible. Ver Angrist y Pischke (2009) para una extensa discusión sobre el tema.

A modo de ejemplo extremo,  $y$  podría ser la cantidad de lluvia en un día en un barrio, y  $x$  la proporción de personas que llevan un paraguas. Siempre es

posible descomponer a  $y$  en una parte relacionada con el uso de paraguas y otra ortogonal a esta, y en base a esta conjetura puede ser posible predecir la cantidad de lluvia en base a la intensidad de uso de paraguas, aun cuando resulte claro que los paraguas no causan lluvia. Esta discusión dista de ser trivial e intenta enfatizar las enormes diferencias que hay entre un modelo predictivo y otro que representa una relación causal.

Consecuentemente, el ejercicio de predecir  $y$  en base a  $x$  consiste en estimar  $E(y|x)$  en base a una datos  $(x_i, y_i)$ ,  $i = 1, \dots, n$ . Una estrategia obvia consiste en proponer

$$y_i = x_i' \beta + u_i$$

y usar el estimador de MCO, de modo que

$$\hat{y}_i = x_i' \hat{\beta}.$$

Las enorme diferencia entre la econometría clásica frecuentista y el aprendizaje automático es que la primera considera al modelo  $y_i = x_i' \beta + u_i$  como dado y concentra sus esfuerzos en estimar  $\beta$  de la mejor manera posible, en donde “mejor” hace referencia a un conjunto de propiedades que refieren a la relación entre  $\hat{\beta}$  y  $\beta$ , tales como insesgadez y varianza mínima. Por el contrario, el aprendizaje automático o estadístico ve a  $x_i' \beta$  como una de las posibles configuraciones de una relación genérica  $f(x_i)$ , candidatas a predecir correctamente a  $y$  fuera de la muestra. Es decir, el objetivo consiste en *construir* y estimar un modelo con fines predictivos sin necesariamente pensar que existe tal cosa como un verdadero proceso que genera los datos.

A modo de ejemplo, pensemos que  $x_i$  es un vector de  $K$  regresores. La econometría clásica toma a la dimensión de este vector como dada, proveniente de la teoría, es decir, es esta última la que dice *ex-ante* que estas  $K$  variables explicativas pertenecen al modelo. Un resultado conocido de la econometría es que la eliminación de variables del modelo esta sujeta a un *trade off* entre sesgo

y varianza: un modelo que incluye todas las variables es posiblemente menos sesgado pero mas ineficiente que uno con menos variables. Los intentos de priorizar la insesgadez llevan al analista a preferir modelos “grandes” al costo de sacrificar eficiencia. Aquí es crucial resaltar que los costos de insesgadez o eficiencia se refieren a  $\hat{\beta}$  en relación a  $\beta$ , el actor protagónico de la econometría tradicional.

### 3.2. Complejidad y elección de modelos

Cuando el foco pasa de  $\beta$  a  $y$ , una enorme diferencia introducida por *machine learning* es que quizás cierto sesgo ocurra a costa de una dramática caída en la varianza. En el contexto discutido anteriormente, al no considerar al modelo como dado, la lógica de aprendizaje estadístico se permite eliminar variables siempre y cuando el sesgo inducido por hacerlo se traduzca en una caída significativa en la varianza de modo que esta última mas que compense el sesgo haciendo así caer al ECM.

En este marco un modelo es mas complejo si tiene mas variables y en base a la lógica anterior, un modelo mas complejo es menos sesgado y mas errático. Consecuentemente, si la insesgadez no es un objetivo primordial no hay en *machine learning* una preferencia obvia por la complejidad, sino que, todo lo contrario, se trata de elegir una complejidad *óptima*. Una parte sustancial del aprendizaje automático (y de la ciencia de la computación) se dedica a generalizar estas ideas de complejidad y optimalidad.

Volviendo al ejemplo simple en donde complejidad es número de variables explicativas, el desafío es proponer una forma de medir la *performance* predictiva de un modelo y en base a ella poner a competir modelos alternativos que difieren en su complejidad, de modo de elegir el mejor.



Como resaltásemos anteriormente, el desafío de *machine learning* es de predicción fuera de la muestra. Herramientas clásicas como el  $R^2$  no funcionan porque miden la *performance* predictiva dentro de la muestra. A modo de ejemplo, un resultado conocido es que el  $R^2$  es no decreciente en la cantidad de regresores (complejidad), de modo que trivialmente siempre el modelo elegido sería el que incorpora mas variables. Interesantemente, viejas estrategias como el  $R^2$  *ajustado* o los criterios de Schwartz o Akaike (populares en series temporales) van en la dirección correcta, en el sentido de que penalizan tanto la falta de capacidad predictiva como la complejidad.

Un contexto ideal en donde esta cuestión puede ser resuelta es si existiese un grupo de datos con los cuales realizar la construcción del modelo y otro para evaluarlo. A la primera muestra se la llama *muestra de entrenamiento* y a la segunda *de evaluación*. En esta circunstancia la práctica consistiría en construir un modelo en base a la muestra de entrenamiento de modo de maximizar su *performance* predictiva en los datos de evaluación.

Hay dos problemas con esta lógica. En primer lugar, no resulta obvio que exista una partición de datos que deban ser usados para estimar y otros para evaluar. En segundo lugar, dejar de lado información implica un costo de eficiencia, es decir, la capacidad predictiva debería mejorar cuanto mas información se utilice, de modo que reservar un subconjunto de datos a fines de evaluar el modelo implica un costo de eficiencia.

### 3.3. Validación cruzada

Una solución inteligente a este problema es la técnica de *validación cruzada*, que consiste en lo siguiente

1. Partir la muestra original en  $J$  grupos disjuntos.
2. Elegir uno de estos grupos y reservarlo como muestra de evaluación. Estimar el modelo con los datos de los  $J - 1$  grupos restantes.
3. Evaluar la capacidad predictiva con los datos del grupo dejado fuera de la estimación.
4. Repetir  $J$  veces el procedimiento reemplazando en cada iteración el grupo de evaluación.

En cuanto a la elección de  $J$  existe amplio consenso empírico en usar  $J = 5$  o  $J = 10$ . Es importante observar que una vez terminado el algoritmo, todas las observaciones fueron usadas tanto para estimar el modelo como para evaluarlo, justamente porque el método “rota” los roles en cada paso.

La estimación del error predictivo por validación cruzada es

$$CV(\hat{f}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_-(x_i))^2$$

en donde  $\hat{y}_-(x_i)$  es la predicción de  $y_i$  en base al modelo estimado cuando la  $i$ -ésima observación no fue utilizada para estimar.

Entonces, una estrategia posible para elegir un modelo sería estimar todos los modelos posibles y elegir aquel que tiene menor error predictivo por validación cruzada.

Salvo casos particulares, la cantidad de modelos posibles a estimar es extremadamente elevada. Por ejemplo, con  $K = 10$  la cantidad de modelos posibles es 3.628.800, que habría que estimar  $J$  veces a fines de computar el error por validación cruzada.  $K = 10$  es una magnitud ínfima en relación al estándar en bases de datos de *big data*. A modo de ejemplo, en una base de datos de información genética de individuos, la cantidad de variables explicativas se cuentan de a *millones*.

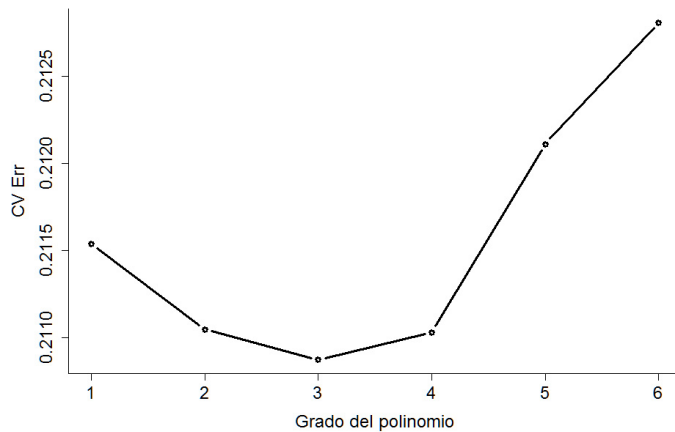
Un ejemplo simple en donde es relevante computar todos los modelos posibles es cuando se plantea una relación polinómica entre  $y$  y  $x$ , es decir

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + u.$$

En este caso la complejidad del modelo se refiere al máximo grado del polinomio. Consecuentemente, una estrategia simple para elegir el grado adecuado es estimar el modelo con una cantidad creciente de potencias de  $x$ , computar para cada uno de ellos el error estándar por validación cruzada, y luego elegir el modelo que lo minimiza.

---

*Ejemplo: Validación cruzada.* El gráfico muestra la estimación del error predictivo por validación cruzada para una sucesión de modelos de regresión en donde el grado máximo del polinomio de la variable explicativa aumenta de a uno. Claramente el mínimo se corresponde con un polinomio de grado 3, a partir del cual si bien el ajuste dentro de la muestra es no decreciente, la capacidad predictiva fuera de la muestra empeora.



---

Lamentablemente en la mayoría de las situaciones prácticas resulta impráctico estimar todas los modelos posibles. Algunas estrategias clásicas consisten en estimar modelos incorporando progresivamente variables o, al revés, comenzando con un modelo genérico e ir reduciendo variables.

Afortunadamente, existen métodos modernos de llevar a cabo esta tarea, como el LASSO, discutido a continuación.

### 3.4. Regularización: LASSO y ridge

Una estrategia muy popular en aprendizaje automático es el *LASSO* introducido por Robert Tibshirani (1996). Esta técnica resuelve en un solo paso el problema de estimar un modelo y elegir que variables retener (reducir su complejidad), en forma similar a la práctica informal (y peligrosa) de estimar un modelo y luego eliminar variables insignificantes. El *LASSO* provee una forma ordenada y eficiente de realizar esta tarea con el propósito de estimar un modelo predictivo con menor ECM fuera de la muestra que uno que incorpora todas las variables explicativas.

El *LASSO* se basa en resolver el siguiente problema de optimización. Para un valor dado  $\lambda \geq 0$ , minimizar

$$L(\hat{\beta}) = \sum_{i=1}^n (y_i - x_i' \hat{\beta})^2 + \lambda \sum_{s=2}^K |\hat{\beta}_s|,$$

con respecto a  $\hat{\beta}$ , en donde  $x_i$  es un vector de  $K$  variables explicativas con su primer elemento igual a uno, de modo que  $\beta_1$  es el intercepto del modelo.

Si  $\lambda = 0$  la solución a este problema esta dada por el estimador MCO. En el otro extremo, cuando  $\lambda = \infty$  la solución consiste en fijar todos los coeficientes  $\hat{\beta}_s$  en cero.

Consecuentemente cualquier valor intermedio  $0 < \lambda < \infty$  produce una estimación  $\hat{\beta}$  que esta contenida entre la mínimo cuadrática y 0. Es decir, para cualquier  $\lambda$  intermedio el *LASSO* es “sesgado” hacia cero en relación al estimador MCO.

¿Cuál es la ganancia de introducir este sesgo? Las razones son dos. En primer lugar, si  $\lambda$  es elegido apropiadamente, el aumento en el sesgo puede ser más que compensado con una drástica caída en la varianza. En segundo lugar, la estructura no diferenciable del problema de optimización de LASSO implica “soluciones de esquina” de modo que ciertos coeficientes son igualados a cero. Es decir, el LASSO “elige” que variables entran al modelo y cuáles no.

El procedimiento por el cual opera este mecanismo es particularmente afín al tipo de razonamiento de los economistas, y vale la pena explorarlo. Sin demasiada pérdida de generalidad, consideremos el caso en donde hay una sola variable explicativa ( $K = 1$ ) y que supondremos estandarizada (con media 0 y varianza 1). El problema de optimización ahora se reduce a minimizar

$$L(\hat{\beta}) = \sum_{i=1}^n (y_i - x_i \hat{\beta})^2 + \lambda |\hat{\beta}|,$$

con respecto a  $\hat{\beta}$ , para un  $\lambda > 0$  dado. Notar que el primer sumando penaliza la falta de ajuste tal como lo hace el método de MCO. El segundo sumando penaliza el mero hecho de que  $\hat{\beta}$  sea no nulo. Esto introduce una suerte de *trade off* en la elección de  $\hat{\beta}$  el primero sumando tiende a elegirlo libremente a fines de minimizar el error agregado, pero el segundo tiende a llevarlo a cero.

La función objetivo de LASSO no es diferenciable lo cual complica discutir el estimador resultante en términos sencillos. Así y todo, es posible caracterizar a la solución a través de una intuición simple. El problema puede ser visto de la siguiente forma. El investigador parte de  $\hat{\beta} = 0$ , y procede a cambiar esta estimación inicial si los beneficios de hacerlo más que compensan los costos.

Es fácil ver que el estimador MCO ( $\hat{\beta}$ ) (correspondiente a  $\lambda = 0$ ) es igual a

$$\hat{\beta}_{mco} = \sum_{i=1}^n y_i x_i$$

ya que las  $x_i$ 's fueron estandarizadas (y por lo tanto  $\sum_{i=1}^n x_i^2 = 1$ ). Supongamos momentáneamente que  $\hat{\beta}_{mco} > 0$ . Consecuentemente, si empezásemos con  $\hat{\beta} =$

0 y  $\lambda$  fuese 0 inmediatamente querríamos llevar  $\hat{\beta}$  a  $\hat{\beta}_{mco}$ . Cuando  $\lambda > 0$ , el beneficio marginal de mover  $\hat{\beta}$  de cero a la derecha (en la dirección de  $\hat{\beta}_{mco}$ ) es

$$\left. \frac{\partial L(\hat{\beta})}{\partial \hat{\beta}} \right|_{\hat{\beta}=0} = \left( -2 \sum (y_i - x_i \hat{\beta}) x_i + \lambda \right) \Big|_{\hat{\beta}=0} = -2 \hat{\beta}_{mco} + \lambda.$$

Es relevante remarcar que la anterior es en realidad una *derivada direccional*, desde  $\hat{\beta} = 0$  a la derecha (la función objetivo no es diferenciable en  $\hat{\beta} = 0$ ). Entonces, partiendo de  $\hat{\beta} = 0$ , el beneficio marginal de moverlo a la derecha es  $2\hat{\beta}_{mco}$  (la caída en la suma de los errores al cuadrado) mientras que el costo de hacerlo es  $\lambda$ . Y aquí es donde conviene razonar como economistas: la función de “beneficios”  $\left( \sum (y_i - x_i \hat{\beta})^2 \right)$  es decreciente y estrictamente cóncava en 0 (porque hemos supuesto que tiene su minimizador a la derecha de 0). Por otro lado, la función de penalidad es lineal ( $\lambda \hat{\beta}$ ) con pendiente  $\lambda$ . Consecuentemente, si en  $\hat{\beta} = 0$  los costos marginales de moverlo a la derecha ( $\lambda$ ) superan a los beneficios ( $2\hat{\beta}_{mco}$ ), llevarlo mas aún a la derecha solo aumentara la diferencia entre los primeros y los segundos ya que costo marginal constante y el beneficio marginal decreciente. Entonces, si en  $\hat{\beta} = 0$

$$-2 \hat{\beta}_{mco} > \lambda,$$

LASSO elige dejar  $\hat{\beta}$  en cero: la variable “no entra” en el modelo estimado. La intuición es clara: LASSO ofrece una beneficio marginal decreciente y una penalización marginal constante por mover  $\hat{\beta}$  de cero, de modo que si la primera supera a la segunda, convendrá dejar  $\hat{\beta} = 0$ .

Mas aún, en este ejemplo simple LASSO elige  $\hat{\beta} = 0$  si  $\hat{\beta}_{mco}$  no es lo suficientemente grande. Es en este sentido que LASSO formaliza la idea de eliminar variables que no son lo suficientemente significativas.

Afortunadamente este caso simple y las ideas que conlleva son generalizables. En primer lugar, cuando  $\hat{\beta}_{mco} < 0$  el argumento funciona exactamente igual pero en forma simétrica (el problema ahora es mover  $\hat{\beta}$  desde

cero a la izquierda). En segundo lugar, la idea de “dejar en cero los coeficientes de variables insignificantes” se aplica de la misma manera al caso en donde hay  $K > 1$  regresores. Ver Hastie et al. (2015) para un tratamiento detallado (y técnico).

Resta por resolver el problema de elección óptima de  $\lambda$ , que hasta ahora hemos tratado como dado. Una estrategia comúnmente utilizada es apelar al método de validación cruzada, que en este caso consiste en: 1) partir de una grilla lo suficientemente grande y fina de valores posibles para  $\lambda$ , 2) estimar todos los modelos posibles para todos los valores alternativos de  $\lambda$  fijados en el paso anterior y computar el ECM por validación cruzada para cada uno de ellos, 3) elegir el  $\lambda$  correspondiente al modelo con menor ECM.

Dada la naturaleza no diferenciable de la función objetivo de LASSO, resulta complejo observar en que sentido una elección inteligente de  $\lambda$  redundante en un ECM menor que el de mínimo cuadrático.

Una forma simple de apreciar esta ganancia es considerar una variante de LASSO, conocida como *ridge*, que se basa en minimizar

$$L(\hat{\beta}_r) = \sum_{i=1}^n (y_i - x_i' \hat{\beta}_r)^2 + \lambda \sum_{s=2}^K \hat{\beta}_{rs}^2.$$

El único cambio en relación a LASSO es que *ridge* penaliza el cuadrado de los coeficientes en vez del valor absoluto. Este cambio aparentemente menor tiene implicancias profundas para el algoritmo. En particular, y a diferencia de LASSO, la función objetivo es ahora diferenciable y el estimador correspondiente puede ser fácilmente derivado en base a técnicas de cálculo diferencial.

En particular, para el caso de  $K = 1$  y  $x_i$  estandarizada es muy fácil mostrar que

$$\hat{\beta}_r = \frac{\hat{\beta}_{mco}}{1 + \lambda}.$$

Notar que en forma análoga al caso de LASSO, el estimador *ridge* es del de MCO “corrido hacia cero”.

Para este caso simple es fácil mostrar que  $E(\hat{\beta}_r) = \beta/(1 + \lambda)$  y  $V(\hat{\beta}_r) = \sigma^2/(1 + \lambda)^2$ . Recordar que estos valores para MCO son  $\beta$  y  $\sigma^2$  respectivamente, ya que es insesgado. Consecuentemente *ridge* es sesgado pero tiene menor varianza. El ECM (la suma del sesgo al cuadrado y la varianza) de *ridge* es

$$ECM(\hat{\beta}) = \left[ \beta - \frac{\beta}{1 + \lambda} \right]^2 + \frac{\sigma^2}{(1 + \lambda)^2}$$

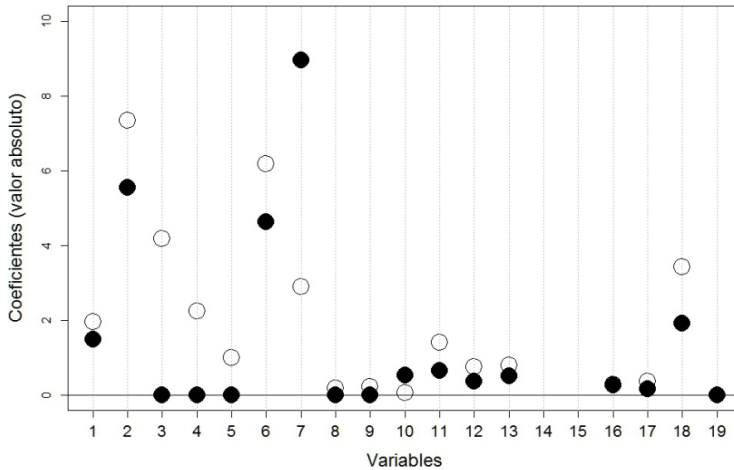
mientras que el ECM para  $\hat{\beta}_{mco}$  es naturalmente  $\sigma^2$ . Un resultado crucial (y de simple verificación algebraica) es que es suficiente que  $\lambda < 2\sigma^2/\beta^2$  para que  $ECM(\hat{\beta}_r) < ECM(\hat{\beta}_{mco})$ . Este resultado es crucial ya que dice que para todo  $\beta$  y para todo  $\sigma^2$  siempre es posible encontrar un valor de  $\lambda$  tal que *ridge* es mejor que MCO en error cuadrático medio. Este es el sentido en el que *ridge* mejora la *performance* predictiva de MCO introduciendo un sesgo que es mas que compensado por una caída en la varianza.

Este resultado es generalizable al caso de  $K > 1$  regresores, el cual provee una aproximación al resultado esperable para el LASSO (para el cual no es posible derivar analíticamente un resultado similar, como aclarásemos anteriormente), en el sentido de que ambos métodos logran bajar el ECM introduciendo un sesgo que permite una dramática caída en la varianza.

Tanto LASSO como *ridge* atacan el *trade off* entre sesgo y varianza a través de un proceso de *regularización*, captado en ambos casos por el sumando adicional que se agrega a la función de perdida (i.e.,  $\lambda \sum |\beta_s|$  para LASSO). Este término penaliza la incorporación innecesaria de variables en el modelo, y en este sentido “regulariza” la complejidad del modelo, estableciendo una suerte de preferencia ex-ante por modelos simples (con menos variables, en este caso).



*Ejemplo: MCO vs. LASSO.* El gráfico muestra resultados para un modelo de predicción de salarios de jugadores de béisbol (ver James et al. (2014) para una discusión de esta base de datos) sobre la base de 19 variables explicativas. El eje horizontal contiene a cada una de las variables explicativas (las variables 14 y 15 se encuentran fuera del rango). Los círculos sin relleno son los valores absolutos de la estimación por MCO de cada coeficiente. Los círculos rellenos son las estimaciones por LASSO. Notar que a) las estimaciones por LASSO son en general menores en valor absoluto que las de MCO, b) LASSO elimina 6 variables, llevando sus coeficientes exactamente a cero.



## 4. Clasificación

### 4.1. Riesgo, probabilidades y clasificación

Cuando la variable objetivo  $y$  es binaria, se dice que el problema de predicción es uno de *clasificación*. Sin pérdida de generalidad supongamos que  $y$  toma solo valores  $\{0, 1\}$  en donde 1 denota pertenencia a un grupo y 0 lo contrario. Por ejemplo, en una muestra de  $n$  personas,  $y_i = 1$  podría corresponder a las que habiendo solicitado un crédito lo pagaron y  $y_i = 0$  a las que no. El problema consiste en predecir  $y$  en base a un vector  $x$  de características observables. En el ejemplo anterior  $x$  podrían ser características de las personas como su status laboral, su ingreso, educación, etc.

Denotemos con  $\hat{y}$  a la predicción de  $y$ . Naturalmente, si  $y$  toma valores 0 o 1,  $\hat{y}$  también lo hará. Consecuentemente el problema consiste en elegir si  $\hat{y} = 1$ , o  $\hat{y} = 0$ . Llamemos  $p \equiv Pr(y = 1|x)$  a la probabilidad de que  $y$  sea 1 dadas las características  $x$ . Obviamente  $1 - p$  es la probabilidad de que  $y$  sea 0 dado  $x$ .

Denotemos con  $\lambda(i, j)$  a la “penalidad” de haber elegido  $\hat{y} = i$  cuando  $y = j$ , para  $i = \{0, 1\}$  y  $j = \{0, 1\}$ . El *riesgo esperado* de realizar la predicción  $i$  es igual a

$$R(i) = (1 - p) \lambda(i, 0) + p \lambda(i, 1), \quad i = 0, 1,$$

es decir, el riesgo esperado de predecir  $y = i$  es el promedio ponderado de las penalidades asociadas a los dos valores que puede tomar  $y$  (0 con probabilidad  $1-p$  y 1 con probabilidad  $p$ ).

Consecuentemente, si  $R(1) < R(0)$  sera conveniente predecir  $\hat{y} = 1$ , y  $\hat{y} = 0$  en caso contrario. Una función de pérdida simple es

$$\lambda(i, j) = 1[i \neq j],$$

en donde  $1[i \neq j]$  es una función indicadora que toma valor 1 si  $i \neq j$  y 0 cuando  $i = j$ . Es decir, esta función de penalidad “castiga” con 1 cuando  $\hat{y}$  no coincide con  $y$ , y con 0 cuando lo hace. Reemplazando en la función de riesgo esperado definida anteriormente,  $R(0) = p$  y  $R(1) = 1 - p$ . Es fácil verificar que en este caso,  $R(1) < R(0)$  siempre que  $p > 0,5$ .

Entonces, cuando la penalización es del tipo binario (0 bajo acierto y 1 cuando no), a los efectos de minimizar el riesgo esperado la regla sugiere predecir la acción mas probable. A esta regla de clasificación se la llama *clasificador de Bayes*.

Consecuentemente, bajo la estrategia de clasificación de Bayes, hemos reducido el problema de clasificación al siguiente procedimiento

1. Construir  $p$  (la probabilidad de que  $y = 1$  dado  $x$ ).
2. Predecir  $\hat{y} = 1$  si  $p > 0,5$ , y  $\hat{y} = 0$  en caso contrario.

Es decir, en este marco el problema de clasificación pasa primero por predecir probabilidades y luego tomar decisiones de clasificación en base a estas probabilidades predichas.

## 4.2. Regresión logística y alternativas

Un popular modelo para estas probabilidades es el *logit*, o modelo de *regresión logística*,

$$p = F(x'\beta),$$

con

$$F(z) = \frac{\exp(z)}{1 + \exp(z)}.$$

Es fácil verificar que  $F(z)$  es monótona y creciente en  $z$  y que toma valores estrictamente en el intervalo  $(0, 1)$ , de modo que para cualquier valor de  $x$  y  $\beta$ ,  $F(z)$  siempre arroja valores compatibles con los de una probabilidad. El vector de parámetros  $\beta$  puede ser estimado consistentemente por el método de máxima verosimilitud en base a una muestra idéntica e independientemente distribuida  $(x_i, y_i)$ ,  $i = 1, \dots, n$ .

El modelo *probit* es una alternativa popular en econometría, que curiosamente recibe muy poca atención en la literatura de *machine learning* y corresponde a

$$F(z) = \int_{-\infty}^x \phi(z) dz,$$

en donde  $\phi(z)$  es la función de densidad para una variable aleatoria normal estándar. Una tercera alternativa es el *modelo lineal de probabilidad*, (*MLP*)

$$F(z) = z,$$

de modo que  $p = x'\beta$ . Este modelo no es otra cosa que el modelo de regresión lineal aplicado sobre una variable dependiente binaria. Claramente, el rango de variación de  $F(z)$  no esta contenido en  $(0, 1)$ , lo que en la práctica puede dar lugar a predicciones de probabilidades menores que cero o mayores que uno, argumento usado en la práctica de *machine learning* para evitarlo. Así y todo, autores como Angrist y Pischke (2009) han argumentado vehementemente que en la práctica econométrica esto es muy raro que ocurra, señalando también que las diferencias empíricas entre el MLP y estrategias no lineales como *logit* y *probit* son mínimas en la práctica. La ventajas del MLP es que su tratamiento ocurre dentro del marco de mínimos cuadrados ordinarios, lo que conduce a rápidas extensiones como modelos para paneles o variables instrumentales, de mas compleja implementación (cuando no inexistente) para el caso no lineal como *probit* o *logit*. Es decir, la práctica econométrica reciente es mucho mas permisiva con el MLP, a diferencia de la de *machine learning*, que lo ignora por completo.

### 4.3. Categorías múltiples

La extensión al caso de mas de dos categorías es delicada. A modo de ejemplo, el problema podría consistir en predecir si un alumno puede ser clasificado como regular, bueno o excelente, o si una persona prefiere música, literatura, cine o teatro como su esparcimiento favorito entre esas cuatro categorías. Estos dos problemas tienen en común que intentan clasificar persona en mas de dos categorías (tres en el primer ejemplo y cuatro en el segundo). Así y todo difieren en un hecho importante: en el primer caso las categorías son ordenables (regular, bueno, excelente) y en el segundo no.

El tipo de modelo usado para clasificar en mas de dos categorías depende de esta distinción. Una estrategia simple para el caso no necesariamente ordenable es el modelo *logístico multiple*, que opera en forma muy similar al *logit*. Es importante remarcar que este modelo implica algunos supuestos fuertes en términos del tipo de relación entre las distintas alternativas (el así llamado supuesto de “independencia de alternativas irrelevantes”) lo que limita su uso. Otra alternativa genérica es el modelo de *análisis discriminante*. Este modelo es fácil de generalizar al caso de mas de dos categorías, pero descansa en supuestos fuertes acerca de la distribución del vector  $x$ . En particular, presupone que  $x$  tiene distribución normal multivariada, lo que puede ser restrictivo en muchas situaciones prácticas, al punto de ser inutilizable cuando  $x$  incluye variables binarias.

#### 4.4. Árboles

Una tecnología flexible para modelos de regresión y clasificación es la de *arboles*, conocida como CART (Classification and Regression Trees). Explica su popularidad el hecho de que en sus versiones mas simples tiene considerables ventajas expositivas.

Supongamos que la variable a predecir ( $y$ ) es binaria y que solo hay dos variables explicativas,  $x_1$  y  $x_2$ . Los datos  $(x_{1i}, x_{2i})$ ,  $i = 1, \dots, n$  pueden ser ubicados en un plano, naturalmente de dimensión dos.

El primer objetivo de CART es partir este plano en dos semiplanos. Si bien existen infinitas formas de hacerlo, CART se concentra solo en un tipo de partición: en base a *una sola* de las variables explicativas. A modo de ejemplo, supongamos que  $y$  es un indicador binario de si una persona que vive con sus padres esta empleada, y que  $x_1$  es su edad y  $x_2$  es el ingreso de sus padres. Una partición posible de los datos podría ser “menores o mayores de 24 años” (en base a la edad) o “cuyos padres ganan mas o menos que 10.000 pesos por mes” (en base al ingreso de sus padres). Dada una partición, CART procede a predecir por mayoría. A modo de ejemplo, si la partición es la del primer ejemplo mencionado, si dentro del grupo de personas menores a 24 años la proporción de jóvenes empleados es 0.4, CART predice que nadie en este grupo empleado, usando la regla de Bayes mencionada anteriormente, que sugiere predecir 0 si  $p < 0,5$  y 0 en caso contrario.

Si este fuese el problema, el objetivo consiste en elegir la mejor partición posible. A tal efecto CART divide a cada variable en una grilla fina y procede a considerar *todas* las particiones posibles, computar para cada partición el error de predicción y elegir la partición que mejor predice. Nuevamente, la evaluación de la capacidad predictiva fuera de la muestra puede realizarse a través de una técnica como validación cruzada.

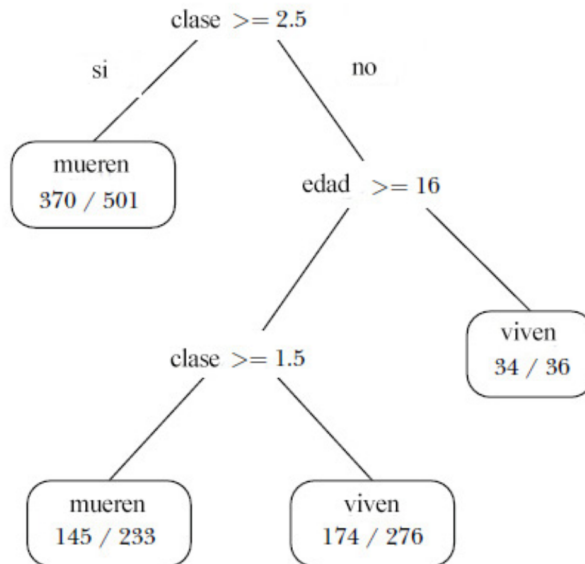
Una vez partido el espacio inicial en dos planos. El algoritmo procede a partir nuevamente estos dos planos en otros dos planos, usando la misma lógica. Es fácil ver que la idea es generalizable a cualquier cantidad de variables explicativas. El algoritmo CART, entonces, hace la siguiente tarea

1. Partir los datos en dos grupos en base a una sola variable explicativa, eligiendo la que predice mejor de entre todas las particiones posibles. La “mejor” se refiere a la que tiene mejor capacidad predictiva.
2. Continuar el proceso de partición anterior dentro de cada una de las particiones obtenidas en el paso anterior.

Resulta claro que este proceso de partición recursiva puede continuar hasta el extremo en donde hay tantas particiones como observaciones. Es decir, el procedimiento empieza con un solo grupo (todas las observaciones) y parte sucesivamente en “ramas” hasta que, en el límite, las ramas jerárquicamente mas debajo contienen una sola observación. Bajo esta estructura, la idea de complejidad se refiere a mas ramas, y en forma análoga a la discusión de la sección 3.2, mayor complejidad “sobreajusta” los datos dentro de la muestra pero no necesariamente fuera de ella. Es decir, en el extremo en el que los datos son partidos en  $n$  ramas (una por observación), el árbol predice perfectamente las observaciones disponibles, pero su *performance* puede ser muy mala fuera de la muestra.

En la tecnología de árboles el análogo de eliminar variables del modelo es eliminar ramificaciones. Es decir, el problema de complejidad óptima pasa por decidir cuando parar de subdividir las ramas, de modo de minimizar el ECM. A la acción de detener el proceso de sub ramificación se la llama (apropiadamente) “poda”. Un ejercicio común es ajustar un árbol con muchas ramificaciones y proceder a “podarlo” (agrupar subdivisiones) de modo de dejar solo unas pocas ramas activas, que predigan más apropiadamente. Una parte relevante de esta literatura se dedica a examinar cual es un mecanismo apropiado para “podar” árboles.

*Ejemplo: CART y el hundimiento del Titanic (tomado de Varian, 2013).* El gráfico muestra resultados de un modelo para predecir que pasajeros murieron en el accidente del Titanic. La variable “clase” se refiere a la categoría en la que viajaban los pasajeros, cuanto mas alta peor (primera, segunda, etc.). El primer nodo predice que si la clase es mayor que 2.5 los pasajeros fallecen (la probabilidad estimada en esta “partición” es  $370/501 = 0,74 > 0,5$ ). En el grupo de las clases mejores es posteriormente partido en grupos de edad, prediciendo que los mayores a 16 años sobreviven. En el grupo restante se predice que los que mueren son los de la peor clase. El factor mas relevante para la predicción es la categoría (clase) y el segundo es la edad.





Este ejemplo simple ilustra varias ventajas de CART como estrategia de clasificación. En primer lugar, las variables más importantes aparecen más arriba en el árbol, resultante de que fueron las primeras elegidas por el algoritmo para particionar en forma óptima. En segundo lugar, CART provee una forma muy didáctica de “explicar” jerárquicamente el proceso de predicción, similar a muchos mecanismos decisorios comúnmente utilizados. En el caso del Titanic, el proceso primero pregunta en qué clase viajaba la persona e inmediatamente ofrece una predicción si responde que en una clase mala, y así sucesivamente. En tercer lugar, notar el carácter marcadamente no lineal de CART. A modo de ejemplo, en el caso de Titanic la variable “edad” solo opera si la persona viajaba en una clase relativamente buena. Es decir, la edad *interactúa* con la clase. CART puede ser visto como una forma conveniente de modelar estas interacciones.

CART provee un muy buen ejemplo del tipo de ganancia que la masividad de datos ofrece a las estrategias de *machine learning*. Cada partición deja dentro del grupo cada vez menos datos. Es decir, ramificar el árbol implica ganar en sesgo pero perder en varianza. Cuando el objetivo es predictivo, el método explota la masividad de datos (locales) para reducir el ECM estimando un modelo más complejo y preciso. Es decir, la existencia de mucha información permite construir un modelo más fino y que predice mejor. Esta es una importante diferencia con la visión clásica de la estadística, en donde el aumento en la información solo impacta en el ECM a través de la reducción de la varianza y no necesariamente a través de la reformulación del modelo. En general, los modelos de *machine learning* son altamente no lineales (como CART) y no paramétricos, por lo cual son muy demandantes en términos de datos, lo que explica las enormes ventajas de la existencia de *big data* en este campo.

Lamentablemente las ventajas expositivas de CART no se traducen en ganancias predictivas en forma simple. Un área activa de investigación son

variaciones de CART que aleatorizan el paso predictivo (como *random forests*) con una notable mejora en la *performance* predictiva, al precio de sacrificar simplicidad expositiva.

## 5. Aprendizaje no supervisado: clusters y dimensionalidad

Todas las situaciones discutidas en la sección anterior son de *aprendizaje supervisado*, es decir, tanto en el problema de regresión como en el de clasificación, es posible observar y para un conjunto de datos, de modo que el problema consiste en “entrenar” al modelo en base a la información de  $y$  y  $x$  y luego utilizarlo para predecir y en base a los nuevos  $x$ . A modo de ejemplo, si el problema fuese el de clasificar a una persona como merecedora de un crédito, la construcción (entrenamiento) del modelo se basa en observaciones previas de  $y$  (individuos que pagaron o no un crédito y sus características  $x$ ). Una vez construido el modelo, las características de un nuevo individuo (su situación financiera, ingreso, educación, etc.) son introducidas en el mismo, que propone una predicción para esta persona (pagara o no el crédito).

Se habla de *aprendizaje no supervisado* cuando los datos previos no ofrecen información acerca del resultado a predecir. Un ejemplo ayudara a aclarar esta cuestión. Supongamos que se observan datos sobre características sociodemográficas de un conjunto de personas, tales como su edad, educación, características de su vivienda, etc. y que el objetivo consiste en clasificar a los individuos en “pobres” y “no pobres”. De no existir ninguna definición obvia de que significa ser pobre, la tarea de agrupar a los individuos en dos categorías (pobre o no pobre) es un ejercicio “no supervisado”, es decir, debe proceder sin observar el resultado deseado: ser pobre. Abusando de algunas cuestiones

notacional, en aprendizaje automático es como si quisiésemos asignarle a cada persona una predicción de  $y$  observando  $X$  pero nunca  $y$ .

En este breve capítulo introductorio nos referiremos a dos situaciones comunes en aprendizaje no supervisado: al análisis de clusters (al que se refiere el ejemplo utilizado) y el problema de reducción de dimensionalidad.

## 5.1. Clusters

El principal objetivo del análisis de clusters consiste en agrupar observaciones. A modo de ejemplo, supongamos que disponemos de una muestra de las calificaciones en 10 exámenes para 40 alumnos de una escuela, y el objetivo es formar dos grupos de alumnos. Central al objetivo del análisis de clusters (y a la noción de aprendizaje no supervisado) es que la información disponible no incluye ninguna variable que indique a que grupo pertenece cada alumno.

Un *cluster* es un grupo que satisface dos criterios: 1) los individuos dentro de un grupo se parecen entre sí, 2) los individuos entre grupos son esencialmente diferentes. Entonces, la idea de *cluster* se relaciona con este doble criterio de similitud *dentro* del grupo y disimilitud *entre* grupos. Es importante remarcar que así entendida, la noción de *cluster* solo garantiza la existencia de grupos en base a esta noción de similitud/disimilitud sin que necesariamente estos grupos tengan a priori ninguna interpretación adicional. A modo de ejemplo, en base al criterio de similitud/disimilitud tal vez la partición óptima resulte en un grupo de pésimos alumnos y otro con el resto, o también en que uno de los grupos contenta a los alumnos muy buenos y el otro al resto. El objetivo de un algoritmo de clusterización consiste en encontrar la forma óptima de agrupar datos en base a esta noción de similitud/disimilitud.

A fines de explicitar un algoritmo para este problema, supongamos que  $\mathbf{X}$  es una matriz  $n \times K$  de  $K$  variables para  $n$  observaciones (en nuestro ejemplo, las  $K = 10$  calificaciones para los  $n = 40$  alumnos. Notar que  $x'_i$  es la  $i$ -ésima fila de la matriz  $\mathbf{X}$ , con  $i = 1, \dots, n$ .

El insumo fundamental del análisis de clusters es la *matriz de distancias*,  $D$ , una matriz  $n \times n$  cuyo elemento  $(i, j)$  es igual a

$$D_{i,j} = d(x_i, x_j), \quad i, j = 1, \dots, n,$$

en donde  $d(x_i, x_j)$  es la distancia entre los vectores  $x_i$  y  $x_j$ . Es importante ver que para cualquier noción de distancia  $D$  es una matriz simétrica ( $D_{i,j} = D_{j,i}$ ) y con diagonal principal nula ( $D_{i,i} = 0$ ).

Un aspecto delicado del análisis de *cluster* es la elección de la noción de distancia. Una alternativa popular es la *distancia euclídea*

$$d(x_i, x_j) = \left[ \sum_{s=1}^K (x_{is} - x_{js})^2 \right]^{1/2}.$$

Supongamos que el objetivo es partir las  $n$  observaciones en  $g = 1, \dots, G$  grupos disjuntos. Un *algoritmo de cluster* es un mecanismo  $C(i)$  que asigna cada observación  $i$  a un único *cluster*  $g$

$$C(i) : \{1, \dots, n\} \rightarrow \{1, \dots, G\}.$$

Consideremos la siguiente función de pérdida:

$$W(C) = \frac{1}{2} \sum_{k=1}^K \left[ \sum_{i,j/C(i)=C(j)=g} d(x_i, x_j) \right].$$

Intuitivamente, para una clasificación  $C$  esta función agrega las disimilitudes *dentro* de cada *cluster*. Un importante resultado es el siguiente

$$\begin{aligned}
 T &= \frac{1}{2} \sum_{i,j} d_{ij} \\
 &= \frac{1}{2} \sum_{g=1}^G \left[ \sum_{i,j/C(i)=C(j)=g} d(x_i, x_j) + \sum_{i/C(i)=g, j/C(j) \neq g} d(x_i, x_j) \right] \\
 &= W(C) + B(C)
 \end{aligned}$$

$T$  es la suma de todas las distancias entre todas las observaciones (la disimilitud total). Este resultado dice que  $T$  es la suma de las disimilitudes *dentro* del *cluster* ( $W(C)$ ) y *entre clusters* ( $B(C)$ ). Como la elección de una clusterización no afecta a  $T$ , minimizar  $W(C)$  con respecto a  $C$  implica maximizar  $B(C)$ . Este resultado es crucial, ya que implica que el doble propósito de minimizar la distancia intracluster y maximizar la distancia entre clusters puede ser alcanzado resolviendo solo uno de los problemas de optimización.

Entonces, el problema de elegir la clusterización óptima sera resuelto a través del problema de elegir  $C$  de modo de minimizar  $W(C)$ . Un algoritmo muy popular para minimizar  $W(C)$  es el de *K-medias*, que sin entrar en formalidades, hace lo siguiente

1. Empezar con una clusterización inicial  $C$  y computar para cada *cluster* el vector de promedios de las  $K$  variables.
2. Reasignar los puntos al *cluster* mas cercano en base a la cuadrado de la distancia euclídea con respecto a las medias calculadas anteriormente.
3. Iterar hasta que no haya reasignaciones.

Se trata de un algoritmo muy popular si bien no exento de problemas. Es posible mostrar que el algoritmo converge, pero solo a un mínimo local. Consecuentemente, la clusterización inicial cumple un rol importante. A modo de ejemplo, Stata y *R* eligen  $G$  observaciones al azar y las toma como si fuesen las medias de cada grupo. Adicionalmente, la media es sensible a *outliers*, por lo que algunos investigadores prefieren reemplazarla por la mediana.

Hasta ahora el método presupone que el número de grupos es conocido de antemano. No existe ninguna forma comúnmente aceptada para determinar el número óptimo de grupos. Trivialmente, aumentar el número de grupos reduce la disimilitud dentro del *cluster*, de hecho si  $K = n$ ,  $W(C) = 0$ . Intuitivamente, el estadístico “gap”, propuesto por Tibshirani (2001) se basa en aumentar progresivamente la cantidad de grupos, y detener el proceso cuando la caída en  $W(C)$  es insignificante.

---

*Ejemplo: Clusters y pobreza.* Caruso et al. (2015) usan métodos de clusters para clasificar a los hogares argentinos en “pobres” y “no pobres”, en base a 19 variables indicadoras del bienestar. El estadístico *gap* de Tibshirani sugiere que la partición en dos grupos es apropiada. El hecho de que es imposible rearmar los grupos en base a una sola variable (como el ingreso) sugiere que el bienestar y la pobreza son fenómenos claramente multidimensionales.

---

## 5.2. Reducción de dimensionalidad

Otro interesante problema no supervisado es el de reducción de la dimensionalidad. A modo de ejemplo, supongamos que se dispone de 5 variables socioeconómicas para un grupo de 200 personas. Estas variables se relacionan con el bienestar de una personas, tal como el ingreso, su educación, etc. Una pregunta relevante es si es posible construir una suerte de índice que permita resumir las 5 variables en una sola, que pueda ser utilizada para medir

el bienestar de las personas. Un ejemplo obvio es construir un índice tomando el promedio simple de las 5 variables. El problema consiste en elegir la mejor forma de construir el índice combinando linealmente las variables iniciales.

Formalmente, supongamos que  $\mathbf{X}$  es una matriz  $n \times K$  que contiene a las  $n$  observaciones de las  $K$  variables. Denotemos con  $\Sigma$  a la varianza de  $\mathbf{X}$ . Notar que por definición se trata de una matriz  $K \times K$ . El índice en cuestión, que llamaremos  $h$ , tiene la siguiente forma

$$h = \mathbf{X}\delta$$

en donde  $\delta$  es un vector de  $K$  números reales. El problema entonces consiste en elegir  $\delta$  óptimamente. Notar que

$$V(\mathbf{X}\delta) = \delta'\Sigma\delta$$

La propuesta consiste en elegir  $\delta$  de modo de maximizar  $V(\mathbf{X}\delta)$ . Maximizar la varianza puede resultar extraño dentro de la lógica de la econometría clásica, pero vale recordar que aquí el objetivo no es estimar ningún parámetro sino buscar la forma mas fehaciente de reproducir la variabilidad de  $K$  variables a través de una sola. Vale recordar que el método de mínimos cuadrados hace exactamente lo mismo: elige  $\hat{\beta}$  de modo que minimizando la suma de errores al cuadrado, maximiza la variabilidad de  $\hat{y}$ .

El problema de maximizar  $\delta'\Sigma\delta$  tiene una solución obvia, que consiste en llevar  $\delta$  a infinito. Consecuentemente, la práctica estándar consiste en normalizar  $\delta$ , es decir maximizar  $\delta'\Sigma\delta$  sujeto a que  $\delta'\delta = 1$ . Llamemos  $\delta^*$  a la solución a este problema. El índice resultante sera

$$h^* = \mathbf{X}\delta^*,$$

que es la mejor combinación lineal que representa la variabilidad contenida en  $\mathbf{X}$ .  $h^*$  es el *primer componente principal*.

Un poco de álgebra permite entender con claridad la naturaleza de la solución a este problema. La función de Lagrange para el problema de optimización restringida es

$$L(\delta, \lambda) = \delta' \Sigma \delta + \lambda (1 - \delta' \delta),$$

en donde  $\lambda$  es el parámetro de Lagrange correspondiente a la restricción de normalización de  $\delta$ . Las condiciones de primer orden para un punto estacionario requieren

$$\Sigma \delta = \lambda \delta.$$

Entonces, en el óptimo,  $\delta$  es el autovector correspondiente al autovalor  $\lambda$  de  $\Sigma$ . Premultiplicando por  $\delta'$ :

$$\delta' \Sigma \delta = \lambda,$$

ya que  $\delta' \delta = 1$  bajo la restricción de normalización.

Entonces, a fines de maximizar  $\delta' \Sigma \delta$  debemos elegir  $\lambda$  igual al máximo autovalor de  $\Sigma$  y  $\delta$  igual al autovector correspondiente. Consecuentemente, el problema de encontrar la mejor combinación lineal que reproduce la variabilidad en  $\mathbf{X}$  se reduce a encontrar el mayor autovalor de  $\Sigma$  y su correspondiente autovector.

El vector  $h^*$  es el primer componente principal. El *segundo componente principal* es la solución al siguiente problema: ¿cuál es la mejor combinación lineal (normalizada) de las variables  $\mathbf{X}$  (en el sentido de varianza máxima discutida anteriormente) que es ortogonal al primer componente principal? Intuitivamente, el segundo componente principal es la “segunda mejor” combinación lineal de  $\mathbf{X}$ , basada en información no contenida en el primer componente principal. En base a esta lógica, para una matriz de  $K$  variables es posible encontrar  $K$  componentes principales, cada uno de ellos ortogonales entre sí (por construcción) y en importancia decreciente.

Una interesante propiedad de este procedimiento es que la variabilidad total en  $\mathbf{X}$  es la suma de los  $K$  autovalores de  $V(\mathbf{X}) = \Sigma$ . Por la forma



en la que fueron construidos los componentes principales, el primero de ellos se corresponde con el autovalor mas alto, el segundo con el que le sigue y así sucesivamente. Entonces, en la práctica el objetivo es retener una cantidad mínima de componentes principales que sea capaz de reproducir una proporción considerable de la variabilidad total de  $\mathbf{X}$ . Una regla comúnmente utilizada es retener solo los componentes cuyo autovalor es mayor que uno.

---

*Ejemplo: Indices de pobreza multidimensional.* Gasparini et al. (2013) utilizan métodos de componentes principales para investigar la dimensionalidad relevante del bienestar. Partiendo de datos de 12 indicadores de bienestar, muestran que su variabilidad puede ser reproducida razonablemente por solo tres componentes principales. El primero de ellos (el mas importante) resulta ser una combinación lineal de ingresos y activos de los hogares, sugiriendo que si bien el bienestar es multidimensional, el ingreso cumple un rol importante en representarlo.

---

## **6. Epílogo: la lenta introducción de *big data* en la economía**

“Cuando se tiene un martillo todos los problemas se parecen a un clavo”, dice la así llamada ley del instrumento, enunciada por el filósofo americano Abraham Maslow en relación a la sobreconfianza en una herramienta, ya sea porque el analista invirtió demasiado en ella o porque está de moda. Y si de modas se trata, de ser *big data* un martillo muchas áreas del conocimiento creen nadar en

un océano de clavos, como sugiere la cita de Chris Anderson de la Introducción a este capítulo.

La economía y en general las ciencias sociales parecen ser una excepción a esta tendencia: sólo recientemente las ideas de *big data* han invadido sus dominios. Lo cual es llamativo en el caso de la economía, a la luz de su histórica afinidad para con lo numérico y lo computacional.

Como adelantásemos en la Introducción, *big data* se refiere a la masividad de datos producidos por las interacciones electrónicas asociadas a circunstancias mundanas como el uso de teléfonos celulares, las transacciones *online* o las provenientes de las redes sociales.

La mera idea de “big” remite a muchos datos, pero los desafíos y problemas de esta nueva tecnología se relacionan con otras cuestiones que van más allá de la masividad. En las antípodas del *big data* parece estar el *small data*, idea asociada a la tradición de la estadística clásica, la de las encuestas de opinión, los sondeos de mercado, los análisis de laboratorio y el propio desarrollo de la econometría. A la luz de los -hasta hace poco- elevados costos de relevamiento, el desafío (mayormente de posguerra) consistió en recabar la menor cantidad de datos que permitiese hablar confiadamente de una población de referencia. La respuesta científica a este problema consistió en estructurar los datos, es decir, en indagar la realidad en forma sistemática de modo que unas pocas observaciones contuviesen suficiente información sobre una población de interés.

Las encuestas estructuradas son un claro ejemplo de esta estrategia: un cuidadoso diseño muestral permite aprender acerca de la población a través de pocos datos. A modo de ejemplo, la Encuesta Permanente de Hogares (EPH), elaborada por el Indec argentino, es la principal fuente de información social del país, y cifras cruciales y polémicas como el desempleo o la pobreza se calculan sobre la base de ella. Muchos se sorprenden de que a fin de dar cuenta de los casi cuatro millones de hogares del Gran Buenos Aires sólo se encuesten

3.029, según datos de la última encuesta disponible. Interesantemente la proporción de hogares entrevistados no es muy diferente de la de sangre extraída para un análisis clínico en relación al total que circula por el cuerpo humano.

Consecuentemente la información proveniente de las fuentes clásicas no es simplemente de pocos datos sino de datos cuidadosamente estructurados para representar en forma adecuada una población relevante.

Desde esta perspectiva *big data* no es sólo más datos del tipo de los que ya había sino un diluvio de datos sin estructura. Y justamente sus enormes desafíos se relacionan con suplir esta falta de estructura, y su contraindicación es creer que más es necesariamente mejor. A modo de ejemplo, usando técnicas de posicionamiento geográfico se podría recabar información milimétrica de los usuarios de una autopista (el tipo de autos utilizados, el uso en horarios pico o el número de ocupantes, etc.), pero muy posiblemente a fines de la política pública la información crucial sea la de los que no la usan. *Big data* puede proveer mucho de lo primero y nada de lo segundo, relativizando la importancia de contar con muchos datos.

Y por esta razón los economistas y los científicos sociales tardaron en abrazar la tecnología de *big data*. La irrupción del fenómeno se dio justo cuando el interés metodológico de la profesión estaba puesto en otros problemas. El grueso de la economía empírica de los últimos veinte años se dedicó a evitar la falacia de la correlación, es decir, a no caer en la banalidad de fomentar el uso de paraguas a fin de resolver una sequía sobre la base de la mera observación de que siempre que la gente anda con paraguas llueve. Para eso, la así llamada “revolución de credibilidad” (ver Angrist y Pischke, 2010) en economía priorizó el uso de datos experimentales, o sea, de datos provenientes de experimentos apropiadamente diseñados en donde, por ejemplo, un plan de entrenamiento laboral es administrado al azar a un grupo de individuos y no a otro, de modo que la diferencia de *performances* se deba al entrenamiento y

no a otra cosa, tal como lo haría un médico probando la eficacia de una droga. Esta forma creíble de generar y analizar datos derivó en un plan sistemático de “evaluaciones de impacto”: dotar a las políticas públicas de una evaluación cuasi experimental que permita aislar y juzgar su efectividad.

Este tipo de razonamiento ocupó una buena parte de la literatura académica y de la política pública de los últimos 20 años, priorizando disponer de unos pocos datos apropiadamente obtenidos (preferentemente a través de un experimento) por sobre muchos datos anárquicos. Consecuentemente, el problema que preocupó a los economistas no era que había pocos datos, sino que eran datos malos: se trataba de datos que no provenían de ningún experimento, sino de la mera observación, como los de los paraguas y la lluvia, y como los que promete *big data* copiosamente. Esta reticencia a confiar ciegamente en los datos explica por qué los economistas entraron tarde al juego de la minería de datos.

Así, la economía no se ha quedado afuera del carro de moda, sino que solo en forma reciente le ha perdido el miedo. A modo de ejemplo, el artículo de Cavallo y Rigobon (2016) ilustra sobre la conveniencia de construir índices de precios sobre la base de técnicas de *scraping*, es decir, “chupando” precios de sitios de venta *online*. Los más optimistas creerían que el índice de precios oficial podría ser rápidamente reemplazado por robots computacionales que toman información de Internet. Los más cautos señalan sus limitaciones, empezando porque los métodos de *scraping* se focalizan sólo en alimentos u otros objetos que se pueden comprar *online*, sesgando la muestra. O el hecho de que una buena parte de los movimientos del índice se relaciona con los bienes tarifados, que trivialmente no son captables a través de plataformas *online*, entre otras dificultades.

Otro ejemplo relevante lo constituye el trabajo de Blumenstock et al. (2015), que en base a una pequeña encuesta de bienestar en Ruanda construye un modelo para medir la pobreza sobre la base de la intensidad de uso de teléfonos

celulares. Es decir, la tecnología permite medir espacial y temporalmente la pobreza en un país acuciado por este flagelo, y en donde por sus características geográficas e institucionales resulta impensable una estrategia sistemática similar a la de nuestra EPH.

Sólo el futuro dirá si el diluvio de datos traerá cambios radicales en la forma en la que los economistas perciben la realidad, similares a los aportados por los experimentos y las evaluaciones de impacto en los últimos veinte años. Se trata de proceder con cautela. Subirse al tren de *big data* porque está de moda es tan necio como no hacerlo por la misma razón, y es crucial sopesar sus ventajas y desventajas. En este sentido, el objetivo de este capítulo es motivar al lector a interesarse en esta temática y realizar una evaluación honesta de sus oportunidades y desafíos.

## **7. Bibliografía sugerida**

Naturalmente, la bibliografía sobre este tema es dinámica y creciente cuando no caótica, de modo que cualquier tratamiento (exhaustivo o no) esta prematuramente condenado a una temprana obsolescencia. Nuevamente, adoptaremos un criterio de mínima: un pequeño número de referencias recientes que permitan al interesado interiorizarse en las nuevas ideas de aprendizaje automático partiendo de la estadística clásica y la econometría.

Una introducción moderna y accesible a la econometría clásica es Wooldridge (2015) o Stock y Watson (2014). Un tratamiento más técnico puede hallarse en Davidson y MacKinnon (2003) o Hansen (2017). En el marco de la econometría, una discusión detallada sobre el uso causal de los modelos de regresión es Angrist y Pischke (2009). A la fecha son escasos los textos

que incorporan ideas de *machine learning* en econometría. Una notable y recomendable excepción es Elliott y Timmermann (2016).

Muchas ideas de aprendizaje automático se derivan de la literatura clásica sobre análisis multivariado. Un texto recomendable es Hardle y Simar (2007). Sobre las relaciones entre la estadística clásica y el paradigma de aprendizaje, una excelente referencia es el texto reciente de Efron y Hastie (2016) y el artículo clásico y premonitorio de Breiman (2003).

Las distinciones entre *machine learning* y aprendizaje automático son sutiles, tal vez la primera enfatiza cuestiones computacionales y la segunda se deriva de la estadística clásica. Un texto recomendable en la primera tradición es el de Murphy (2012). En el segundo grupo se encuentra el libro de Hastie et al. (2016) y también el de James et al. (2015), de un nivel técnico mas accesible y con numerosos ejemplos.

Las cuestiones computacionales son centrales a esta literatura. R es un lenguaje de programación flexible y gratuito con una extendida red de usuarios. Se encuentra disponible en <https://cran.r-project.org>, en donde se puede hallar una interesante colección de documentación tanto para iniciados como para expertos. Una buena introducción a los métodos de aprendizaje en base a R es Kuhn y Johnson (2014) con una gran cantidad de ejemplos empíricos. Python es también una alternativa relevante. Una introducción accesible y con ejemplos es Muller y Guido (2016).

En lo que se refiere al uso de *big data* e ideas de aprendizaje en economía, algunos artículos introductorios muy recomendables son Varian (2014) y Einav y Levin (2014). Un muy buen artículo sobre los usos causales y predictivos de modelos en econometría esta en Kleinberg et al. (2015).

## Referencias

- Anderson, C. (2008). The end of theory. *Wired*, 16(7), 16-07.
- Angrist, J. y Pischke, J. (2009). *Mostly Harmless Econometrics: An Empiricist Companion*. Princeton University Press. Princeton.
- Angrist, J. y Pischke, J. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2): 3-30.
- Bajari, P., Nekipelov, D., Ryan, S. P. y Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, 105(5), 481-485.
- Blumenstock, J., Cadamuro, G. y On, R. (2015). Predicting poverty and wealth from mobile phone metadata, *Science*, 350(6264), 1073-1076.
- Breiman, L. (2003). Statistical modeling: The two cultures. *Quality control and Applied Statistics*, 48(1), 81-82.
- Caruso, G., Sosa Escudero, W. y Svarc, M. (2015). Deprivation and the Dimensionality of Welfare: A Variable-Selection Cluster-Analysis Approach, 2015, with German Caruso and Marcela Svarc. *Review of Income and Wealth*, 61, 4, 702-722.
- Cavallo, A. y Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives*, 30, 2, 151-78.
- Davidson, R. y MacKinnon, J. (2003). *Econometric Theory and Methods*. Oxford University Press, Oxford.
- Einav, L. y Levin, J. (2014). Economics in the age of big data. *Science*, 34-62.
- Efron, B. y Hastie, T. (2016). *Computer Age Statistical Inference*, Cambridge University Press, Cambridge.
- Einav, L., Knoepfle, D., Levin, J., y Sundaresan, N. (2014). Sales taxes and internet commerce. *American Economic Review*, 104(1), 1-26.

- Elliott, G. y Timmermann, *Economic Forecasting*, Princeton University Press, Princeton.
- Friedman, J., Hastie, T., y Tibshirani, R. (2001). *The Elements of Statistical Learning*, Springer, Berlin.
- Gasparini, L., Marchionni, M., Olivieri, S. y Sosa Escudero, W. (2013). Multidimensional Poverty in Latin America and the Caribbean: New Evidence from the Gallup World Poll, *Journal of Economic Inequality*, 11, 195-214.
- Hansen, B. (2017). *Econometrics*, mimeo, University of Wisconsin at Madison.
- Hardle, W. y Simar, L. (2015). *Applied Multivariate Statistical Analysis*, Springer, New York.
- Hastie, T., Tibshirani, R. y Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall, New York.
- James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer. New York.
- Keely, L. C. y Tan, C. M. (2008). Understanding preferences for income redistribution. *Journal of Public Economics*, 92(5), 944-961.
- Kleinberg, J., Ludwig, J., Mullainathan, S. y Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491-495.
- Kuhn, M. y Johnson, K. (2014). *Applied Predictive Modeling*, Springer, New York.
- Muller, A. y Guido, S. (2016), *Introduction to Machine Learning with Python: A Guide for Data Scientists*, O'Reilly Media, New York.
- Murphy, K., (2012). *Machine Learning: a Probabilistic Perspective*, MIT Press, Cambridge.
- Stock, J. y Watson, M., (2015), *Introduction to Econometrics*, Pearson, New York.



- Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological)*. 58, 1, 267-288.
- Tibshirani, R., Walther, G. and Hastie, T. (2001), Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63: 411-423.
- Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3-27.
- Wooldridge, J. (2015). *Introductory Econometrics: a Modern Approach*, South-Western College Publishers, New York.



## **Sobre los autores**

### **Hildegart Ahumada**

Es Licenciada en Economía de la Universidad de La Plata, MPhil in Economics de la Universidad de Oxford y John S. Guggenheim Fellow. Es Miembro Titular y de la Junta Directiva de la Academia Nacional de Ciencias Económicas. Actualmente es profesora e investigadora en la Escuela de Gobierno. Fue Presidente de la AAEP (2000/2002), Supervisor de Investigaciones Económicas del Banco Central de la República Argentina e Investigador visitante en Nuffield College (University of Oxford). Sus principales áreas de investigación cubren una amplia gama de enfoques econométricos (series de tiempo, datos de panel y pronóstico) aplicados al estudio del comportamiento y proyección de variables macroeconómicas y sectoriales con contribuciones más citadas en demanda de dinero, decisiones de consumo, economía oculta y precio de commodities. Actualmente su campo de investigación se focaliza en la econometría del cambio climático y la selección automática de modelos econométricos. Sus publicaciones incluyen artículos en prestigiosas revistas locales e internacionales.

### **M. Florencia Gabrielli**

Es Licenciada en Economía de la Universidad Nacional de Cuyo (UNCuyo), obtuvo una Maestría en Políticas Públicas de la Universidad Torcuato Di Tella en el año 2000 y un Doctorado (PhD) en Economía en The Pennsylvania State University en 2009. Trabajó como analista de investigación en la Gerencia de Investigaciones del Banco Central de la República Argentina durante el período 2000-2004. Actualmente es investigadora del CONICET y es docente de grado y posgrado en la Facultad de Ciencias Económicas de la UNCuyo. Ha publicado varios artículos en revistas científicas nacionales e internacionales. Ha realizado visitas académicas a Australian National University (Australia), Universidad de Chicago (EE.UU.), Universidad de Virginia (EE.UU), Universidad Adolfo Ibañez (Chile), Universidad Alberto Hurtado (Chile) y Universidad de San Andrés (Argentina). Sus principales áreas de interés e investigación son Econometría, Organización Industrial y Microeconomía Aplicada.

### **Marcos Herrera**

Es Licenciado en Economía de la Universidad Nacional de Río Cuarto, Magíster en Economía (Universidad Nacional de Tucumán), Master y Doctor en Economía por la Universidad de Zaragoza (2011), incluyendo posgrado en el Spatial Econometrics Advanced Institute (Università degli Studi di Roma La Sapienza, 2010). Es Investigador Adjunto del CONICET y Vice-Director del IELDE de la Universidad Nacional de Salta. Actualmente es profesor invitado en la Maestría de Economía de la Universidad de San Andrés y del Doctorado de Ciencias Económicas de la Universidad Nacional de Cuyo. Sus publicaciones se encuentran en prestigiosas revistas locales e internacionales. Realizó visitas académicas a diferentes instituciones como: Institute for Studies and Economic Analyses (Italia), Université Nice Sophia-Antipolis (Francia)

y Luxembourg Institute of Socio-Economic Research (Luxemburgo). Su investigación está enfocada en Econometría Espacial teórica-aplicada.

### **Walter Sosa Escudero**

Es Licenciado en Economía de la UBA y Ph.D. en Economía de la Universidad de Illinois (Urbana-Champaign). Es profesor a tiempo completo en la Universidad de San Andrés, en la cual ha dirigido su departamento de Economía y sus programas de maestría y licenciatura. Es profesor titular de Econometría en la UNLP e investigador invitado del CEDLAS de la UNLP. Ha recibido los premios Fulvio Pagani/ARCOR, Bernardo Houssay (Investigador Joven), Konex (2005-2015), el Premio EGSO como el mejor profesor del PhD en economía de la Universidad de Illinois y elegido por la Fundación BGH como uno de los 100 argentinos innovadores. Fue presidente de la AAEP (2012/14). Es Investigador Principal del CONICET y miembro titular de la Academia Nacional de Ciencias Economicas. Su investigación se concentra fundamentalmente en Econometría con aplicaciones a Microeconomía y cuenta con publicaciones en destacadas revistas nacionales e internacionales.



Se terminó de imprimir en el mes de octubre de 2018,  
en la Ediuns, Santiago del Estero 639, Bahía Blanca, Argentina  
Se imprimieron 200 ejemplares







