Econometría II

Trabajo Práctico N°1

Tomás M. Bustos
tomasmbustos@gmail.com
n° registro: 884.781

Lorenzo Perrotta
loren.perrotta@hotmail.com
n° registro: 882.969

Profesor: Luís A. Trajtenberg

Segundo Cuatrimestre de 2018

1. Ejercicio 1 - Variables proxy

Siendo (1.1) el modelo estructural, resultará de interés estimar el efecto causal de $\boldsymbol{X}$ sobre Y.

$$\begin{aligned} Y_i &= E[Y_i|\boldsymbol{X}_i, q_i] + \varepsilon_i \\ E[Y_i|\boldsymbol{X}, q] &= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \gamma q_i \\ E[\varepsilon|\boldsymbol{X}, q] &= 0 \end{aligned} \tag{1.1}$$

La estrategia de identificación elegida para $\beta_1, \beta_2, \ldots, \beta_K$ es el uso de la forma reducida:

$$Y_i = b_0 + b_1 X_{1i} + \cdots + b_K X_{Ki} + \tau Z_i + \varepsilon_i \tag{1.2}$$

siendo Z una variable proxy para la heterogeneidad inobservable q. La estrategia elegida tendrá mayor o menor efectividad dependiendo de cuan "buena" sea la variable proxy para mitigar la inconsistencia generada por la inobservabilidad de q. Siendo (1.3) la proyección lineal de Z_i sobre q_i, la calidad de la variable proxy dependerá de su relación con el conjunto de información $\Omega = \{Y, X_1, \ldots, X_K, q\}$.

$$q_i = \theta_0 + \theta_1 Z_i + r_i \tag{1.3}$$

¿Qué problemas pueden surgir en el proceso de identificación de $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$ ante la presencia de una variable inobservable q correlacionada tanto con Y como con $\boldsymbol{X}$? Suponga que partimos de un modelo como el planteado en (1.2) sin la inclusión de Z.

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{OLS} &= \left[\frac{\sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{X}_i}{N}\right]^{-1} * \frac{\sum_{i=1}^N \boldsymbol{X}_i' Y_i}{N} \\ &= \left[\frac{\sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{X}_i}{N}\right]^{-1} * \frac{\sum_{i=1}^N \boldsymbol{X}_i'(\boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i)}{N} \\ &= \boldsymbol{\beta} + \left[\frac{\sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{X}_i}{N}\right]^{-1} * \frac{\sum_{i=1}^N \boldsymbol{X}_i'\varepsilon_i}{N} \\ Plim\ \hat{\boldsymbol{\beta}}^{OLS} &= \boldsymbol{\beta} + E[\boldsymbol{X'X}]^{-1} * E[\boldsymbol{X'}\varepsilon] \end{aligned}$$

Si se satisfacen las restricciones de identificación del problema, $E[\boldsymbol{X}'\varepsilon] = 0$ y **OLS** generaría estimadores consistentes para $\boldsymbol{\beta}$. Sin embargo, dado que $\varepsilon = \gamma q + v$ y $Cov(q; \boldsymbol{X}) \neq 0$, **OLS** produce estimadores inconsistentes para $\boldsymbol{\beta}$.

$$Plim\ \hat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta} = \ \gamma * E[\boldsymbol{X'X}]^{-1} * E[\boldsymbol{X'}q] \tag{1.4}$$

Los requerimientos para que una variable proxy pueda mitigar la inconsistencia generada por la omisión de variables son:

Redundancia en el modelo estructural Una vez que se controla por $\boldsymbol{X}$ y q, la variable proxy debe ser redundante en sentido de esperanza condicional en (1.1). Cuando se trabaja con un modelo aditivo es suficiente que:

$$Cov(Z, \varepsilon) = 0$$

Redundancia de *X* para explicar *q* una vez controlado por *Z* La variable proxy Z debe tener suficiente poder explicativo sobre q tal que $\boldsymbol{X}$ sea redundante para explicar q.

$$Cov(\boldsymbol{X}, r) = 0$$

Caso 1: Entendiendo que, por definición, (1.5) es una propiedad de la proyección lineal planteada en (1.3) y que la expresión (1.6) expresa la irredundancia de $\boldsymbol{X}$ para explicar q una que se controló por Z, la inclusión de Z en el modelo permitiría estimar consistentemente $\beta_1, \beta_2, \ldots, \beta_K$. La expresión (1.7) no es relevante al problema: sin pérdida de generalidad se puede suponer $E[r_i] = 0$ y solo se vería afectado el intercepto de la ecuación (1.3).

$$Cov(Z_i, r_i) = 0 \tag{1.5}$$
$$Cov(X_{ji}, r_i) = 0 \qquad \forall\, j = 1, 2, \ldots, K \tag{1.6}$$
$$E[r_i] \neq 0 \tag{1.7}$$

Siendo (1.1) el modelo estructural, la utilización de variables proxy nos da acceso a la ecuación estimable (1.8) permitiendo comprender las consecuencias de correr OLS con un modelo como (1.2).

$$Y_i = (\beta_0 + \omega(\theta_0 + \gamma)) + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + (\omega\theta_1) Z_i + (\omega(r_i - \gamma) + \varepsilon_i) \tag{1.8}$$

Por tanto, pueden estimarse consistentemente los parámetros de interés:

$$Plim\ \hat{\beta}_j^{OLS} = \beta_j\,, \quad \forall j = 1, \ldots, K \tag{1.9}$$

Caso 2: Así como en el primer caso, (1.12) es irrelevante y (1.10) es una propiedad de las proyecciones lineales. Sin embargo, (1.11) hace que Z sea un control imperfecto. Esto es, no tiene el suficiente poder explicativo como para mitigar el efecto parcial de X_K sobre q en la ecuación (1.3), implicando que $\hat{\beta}_K^{OLS}$ será un estimador inconsistente para β_K.

$$Cov(Z_i, r_i) = 0 \tag{1.10}$$
$$Cov(X_{ji}, r_i) \neq 0\,, \qquad \forall\, j = K \tag{1.11}$$
$$E[r_i] = 0 \tag{1.12}$$

Siendo q la variable omitida, puede expresarse según:

$$q_i = \theta_0 + \rho_K X_{Ki} + \theta_1 Z_i + r_i \tag{1.13}$$

Esto implica que la forma reducida planteada en (1.2) puede interpretarse en función de los parámetros del modelo estructural según:

$$Y_i = \ (\beta_0 + \gamma\theta_0) + \beta_1 X_{1i} + \cdots + (\beta_K + \gamma\rho_K) X_{Ki} + (\gamma\theta_1) Z_i + (\gamma r_i + \varepsilon_i) \tag{1.14}$$

Así, **OLS** genera un estimador inconsistente para β_K dado que la variable proxy no logra mitigar la fuente de endogeneidad para X_K tendiendo a sub-estimar o sobre-estimar sistemáticamente su efecto sobre Y dependiendo de los signos de γ y ρ_K

$$Plim\ \hat{\boldsymbol{\beta}}_j^{OLS} - \boldsymbol{\beta}_j = \omega * \rho_j\,, \quad \forall\, j = 1, \ldots, K-1 \qquad\qquad Plim\ \hat{\boldsymbol{\beta}}_K^{OLS} - \boldsymbol{\beta}_K = \omega * \rho_K$$

Dado que q está correlacionado con X_K y con Y habrá una parte del efecto medido de X_K que será consecuencia del efecto de q sobre Y y que somos incapaces de separar mediante el uso de una variable proxy imperfecta. Sin embargo, Z permite estimar consistentemente $K-1$ efectos específicos y, en general, su introducción en la ecuación (1.13) hará reducir ρ_K al "limpiarlo" de la parte de q correlacionada con Z y, por tanto, reduciendo su inconsistencia. Por otro lado, su introducción en la ecuación estructural puede ser beneficiosa en la medida en que pueda reducir $V(\varepsilon)$ sin introducir un grado mayor de colinealidad que compense tal reducción.

Caso 3: Si bien en este caso Z tiene el suficiente poder explicativo para "limpiar" la correlación entre q y $\boldsymbol{X}$, sucede que no es redundante en el modelo estructural para explicar Y en sentido condicional, una vez controlado por q.

$$Cov(Z_i, r_i) = 0 \tag{1.15}$$
$$Cov(X_{ji}, r_i) = 0\,, \qquad \forall\, j = 1, \ldots, K \tag{1.16}$$
$$Cov(Z_i, \varepsilon_i) \neq 0 \tag{1.17}$$

Debido a que Z es endógena, no se satisfacen las restricciones de identificación del modelo y, por tanto, **OLS** genera estimadores inconsistentes para todas las variables de interés. Puede notarse en (1.18) que, los regresores son ortogonales al error del modelo pero, debido a (1.17), Z es endógena.

$$Y_i = \ (\beta_0 + \omega\theta_0) + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + (\omega\theta_1) Z_i + (\omega r_i + \varepsilon_i) \tag{1.18}$$

Por lo tanto, para los parámetros de interés:

$$Plim\ \hat{\beta}_j^{OLS} \neq \beta_j\,, \quad \forall j = 1, \ldots, K \tag{1.19}$$

En la práctica, al observar unidades económicas individuales, debe comprenderse que hay muchos factores inobservables que influyen sobre el proceso generador de decisiones. Tal problema está presente en la identificación de los "retornos a la escolaridad formal" en la ecuación de salarios.

$$\begin{aligned} log(wage_i) &= E[log(wage_i)|educ_i, abil_i] + \varepsilon_i \\ E[log(wage_i)|educ_i, abil_i] &= \beta_0 + \beta_1 educ_i + \gamma abil_i \\ E[\varepsilon_i|educ_i, abil_i] &= 0 \end{aligned} \tag{1.20}$$

El modelo estructural está dado por (1.20) y nos interesa estimar consistentemente β_1. Dado que la habilidad individual está correlacionada tanto con $wage$ como con $educ$ y, a su vez, no es observable por el investigador, puede hacerse uso de variables proxy como estrategia de identificación:

$$\begin{aligned} log(wage_i) &= \beta_0 + \beta_1 educ_i + v_i \\ v_i &= \gamma abil_i + \varepsilon_i \\ abil_i &= \theta_0 + \theta_1\, IQ_i + r_i \end{aligned} \tag{1.21}$$

La forma reducida está dada por:

$$log(wage_i) = (\beta_0 + \gamma\theta_0) + \beta_1 educ_i + \gamma\theta_1\, IQ_i + (\gamma r_i + \varepsilon_i) \tag{1.22}$$

Si en (1.21) efectivamente $educ$ es redundante para explicar la habilidad individual una vez que se controló por el resultado del IQ test y este último es redundante en (1.20), entonces se podrá estimar consistentemente β_1. Sin embargo, deben hacerse dos aclaraciones:

1. Lo más probable es que IQ no tenga el suficiente poder explicativo sobre $abil$ como para mitigar el efecto de $educ$. En tal caso, se tendrá una variable proxy imperfecta que, en general, reducirá la inconsistencia del estimador para β_1 bajo **OLS**.

$$Plim\ \hat{\beta}_1^{OLS} \neq \beta_1$$

2. Por otro lado, puede que el modelo en (1.20) esté mal especificado debido a la omisión de otras variables relevantes en el proceso generador de $log(wage)$ que también están correlacionadas con $educ$. En tal caso, estaría confundido nuevamente el efecto específico de los años de escolaridad formal sobre el salario medio por hora. Si estas variables son observables por el investigador, lo más atinado sería incluirlas como controles en el modelo estructural para reducir las fuentes de endogeneidad. Algunas de ellas pueden ser la edad, experiencia, rama de la actividad en la que el individuo se desempeña laboralmente, zona geográfica, entre otros.

2. Ejercicio 2 - Analíticos y empíricos

Ejercicios de Wooldridge (2010)

Ejercicio 5.2

Considerando el siguiente modelo para la salud de un individuo:

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u_1 \quad (2.1)$$

a) ¿Cómo le parece que podría ser que $exercise$ esté correlacionado con el error u_1?

El modelo presentado en (2.1) podría estar mal especificado por varias causas:

(I) La presencia de variables externas al modelo que afectan a la decisión de hacer ejercicio y que, a su vez, hacen a un individuo más saludable, hace que la identificación del efecto específico que tiene el ejercicio sobre la salud quede confundido. La exclusión de variables como el nivel de ingresos, el hecho de si el individuo fuma o posee alguna otra enfermedad, entre otras posibles características del individuo, hace que ocurra un *sesgo por selección* que confundiría una asociación estadística con una mera relación causa-efecto. Esto es, las personas que hacen ejercicio suelen tener características inobservables diferentes a los que no lo hacen que pueden estar correlacionadas con una vida mas saludable.

(II) El problema de simultaneidad entre $health$ y $exercise$ crea otra potencial fuente de endogeneidad. En general, los individuos saludables se autoseleccionan para hacer ejercicio (mientras que puede ser prohibitivo para los que gozan de mala salud) mientras que, en dirección contraria, el ejercicio tiene impacto sobre la salud.

b) Suponga que se pudieran recolectar datos de dos variables adicionales: $disthome$ y $distwork$, que son la distancia a un gimnasio (o club deportivo) desde tu casa o desde el trabajo, respectivamente. Discuta si estas no estarían correlacionadas con u_1.

La inclusión de estas dos nos remite nuevamente al análisis previo. $disthome$ y $distwork$ son potencialmente endógenas debido a la correlación con variables omitidas en el modelo planteado en (2.1). Puede que los gimnasios estén generalmente localizados en barrios de altos y medianos ingresos, variable que identificamos anteriormente como relevante en el proceso generador de $health$.

c) Asuma que $disthome$ y $distwork$ no están correlacionadas con el error u_1, como todas las variables en la ecuación (2.1), a excepción de $exercise$. Escriba la forma reducida para $exercise$ y las condiciones bajo las cuales los parámetros de la ecuación (2.1) están identificados.

Si tanto $disthome$ como $distwork$ son exógenas en (2.1), pueden utilizarse apropiadamente para inducir variabilidad exógena en $exercise$ según (2.2) en tanto se verifique la condición (2.3).

$$\begin{aligned} exercise = \delta_0 + \delta_1 age + \delta_2 weight + \delta_3 height + \delta_4 male + \delta_5 work \\ + \theta_1 disthome + \theta_2 distwork + r \end{aligned} \quad (2.2)$$

$$\theta_i \neq 0 \quad para\ al\ menos\ un\ i = 1, 2 \quad (2.3)$$

Es decir, que el efecto parcial de $disthome$ o $distwork$ debe ser no nulo una vez controlado por el conjunto de variables exógenas (que actúan como instrumentos de si mismos). Si ello sucede, el problema de identificación para $\boldsymbol{\beta}$ tiene solución única y variables instrumentales (IV) provee de estimadores consistentes. Si tanto θ_1 como θ_2 son no nulos,

existen infinitas soluciones para el problema de identificación de β (se genera un sistema compatible indeterminado). En tal caso, resulta deseable utilizar la combinación de instrumentos que maximice la variabilidad exógena de $exercise$ mediante mínimos cuadrados en dos etapas (2SLS).

d) ¿Cómo se pueden probar los supuestos de identificación del punto *c*)?

En el punto *c*) tomamos como dada la exogeneidad de los instrumentos $disthome$ y $distwork$, siendo la única restricción a probar la "relevancia" de los instrumentos. Esto es sencillo con un test de múltiples restricciones de exclusión sobre (2.2) en donde el espacio muestral del estadístico de prueba (2.7) se parte según las hipótesis presentadas en (2.6).

$$\begin{cases} H_0: & \theta_1 = \theta_2 = 0 \\ H_1: & no\ H_0 \end{cases} \tag{2.4}$$

$$F = \frac{[SSR_R - SSR_{NR}]/2}{SSR_{NR}/(N-8)} \sim \mathcal{F}_{2;N-8} \tag{2.5}$$

En donde se comparan las pérdidas de información SSR_R y SSR_{NR} del modelo restricto y no restricto, respectivamente. En caso de ser necesario, podría utilizarse la versión robusta a la heterocedasticidad del test para probar la significatividad conjunta de los instrumentos o, mismo, podría utilizarse el método LM tanto en contextos de homocedasticidad como heterocedasticidad. Puede que sea incluso necesario un test de exclusión individual para cada uno de los instrumentos, si se tuviera la hipótesis de que alguno es irrelevante por si mismo. Nuevamente, si se está en un contexto con heterocedasticidad, se debería implementar el test con la corrección a la matriz de varianzas y covarianzas propuesta por White (1980).

$$\begin{cases} H_0: & \theta_i = 0 \\ H_1: & no\ H_0 \end{cases} \tag{2.6}$$

$$t = \frac{\hat{\theta}_i^{OLS}}{S_{\hat{\theta}_i^{OLS}}} \sim \mathcal{T}Student_{N-8} \tag{2.7}$$

Por último, cabe mencionar, que si bien no puede probarse la exogeneidad de los instrumentos $disthome$ y $distwork$ por la inobservabilidad del error estructural, podría implementarse el test de *restricciones de sobre-identificación* en tanto θ_1 y θ_2 sean conjuntamente no nulos. De todas formas, como se mencionó, el ejercicio da por hecho que los instrumentos están incorrelacionados con el error estructural.

Ejercicio 5.3

Considerando el siguiente modelo para el peso de los recién nacidos:

$$log(bwght) = \beta_0 + \beta_1 male + \beta_2 parity + \beta_3 log(faminc) + \beta_4 packs + u \tag{2.8}$$

a) ¿Por qué podría esperarse que $packs$ esté correlacionada con u?

En el modelo presentado, $packs$ es potencialmente endógena debido a la presencia de variables no incluidas explicitamente en (2.8) que afectan a la decisión de consumo de cigarrillos y, adicionalmente, al peso del recién nacido. Las mujeres fumadoras que sufren de estrés y ansiedad en el embarazo posiblemente sean un problema para este modelo; asimismo las madres fumadoras, en media, pueden tener hábitos alimenticios menos saludables en relación a una madre que no fuma (cafeína, alcohol, alimentos menos nutritivos, salto de comidas, etc.).

b) Suponga que se tienen datos del precio promedio de los cigarrillos en cada uno de los estados donde residen las mujeres, discuta si esta información puede satisfacer las propiedades de un buen instrumento para packs.

En principio, un buen instrumento debe ser tanto "válido", como "relevante". Esto significa que $Cov(cigprice, u) = 0$, es decir, que sea redundante en proceso generador de $log(bwght)$, y que $\theta_1 \neq 0$, siendo θ_1 el efecto parcial de $cigprice$ sobre $packs$ una vez controlado por el otro conjunto de instrumentos. Lo que se debería esperar es que $cigprice$ sea exógena para cada individuo en (2.8) si se toman los controles necesarios. Pues, es probable que el precio de los cigarrillos en cada estado esté correlacionado con distintos factores regionales que afectan la salud de una madre embarazada y, por tanto, el peso del recién nacido, según lo mencionado en Wooldridge (2010, p. 95). Por otro lado, si bien se espera que el precio de los cigarrillos esté correlacionado negativamente con el consumo medio en cada estado, existen argumentos por los cuales la correlación podría ser débil: (*i*) debido a la fuerte agregación de preferencias que representa un precio promedio estatal y (*ii*) debido a la posibilidad de que la demanda de cigarrillos sea más bien inelástica y las variaciones en los precios expliquen poco variaciones en el consumo.

c) Usando los datos en BWGHT.RAW para estimar la ecuación (2.8), estimar primero por **OLS** y luego por **2SLS** donde $cigprice$ es un instrumento para $packs$. Discutir si se encuentra alguna diferencia entre las dos estimaciones.

	OLS	2SLS
male	0.0262**	0.0298
	(0.0101)	(0.0177)
parity	0.0147**	-0.00124
	(0.00566)	(0.0219)
lfaminc	0.0180**	0.0636
	(0.00558)	(0.0569)
packs	-0.0837***	0.797
	(0.0171)	(1.084)
Constant	4.676***	4.468***
	(0.0219)	(0.258)
Observations	1388	1388
R^2	0.035	.

Standard errors in parentheses
* $p < 0{,}05$, ** $p < 0{,}01$, *** $p < 0{,}001$

Cuadro 1: OLS vs. 2SLS
Fuente: *BWGHT.RAW (Wooldridge, 2010)*

En el Cuadro 1 puede notarse que la estimación del efecto parcial de $packs$ sobre $bwght$ difiere significativamente entre ambas estrategias de identificación. Por un lado, la estimación por **OLS** sugiere que un paquete de cigarrillos adicional puede reducir el peso de un recién nacido en un 8.37 %, siendo estadísticamente significativo. Por el otro, sorpresivamente, la estimación del efecto específico mediante **2SLS** no solamente tiene el signo opuesto, sino que un paquete de cigarrillos adicional aumentaría $log(bwght)$ un 79,7 % con un *p-value* 46.2 %. La magnitud de la varianza del estimador $\hat{\beta}^{2SLS}_{packs}$ sugiere que el instrumento utilizado es muy pobre o, en otros términos, no logra inducir suficiente variabilidad exógena sobre $packs$ para la precisa estimación de β_{packs}.

d) Estimar la forma reducida para $packs$. ¿Qué se concluye de la ecuación (2.9) usando $cigprice$ como un instrumento para $packs$? Qué efecto tiene esta conclusión en tu respuesta de la pregunta *c*).

	(1) packs
male	-0.00473 (0.0159)
parity	0.0181* (0.00888)
lfaminc	-0.0526*** (0.00870)
cigprice	0.000777 (0.000776)
Constant	0.137 (0.104)
Observations	1388
R^2	0.030

Standard errors in parentheses
* $p < 0{,}05$, ** $p < 0{,}01$, *** $p < 0{,}001$

Cuadro 2: Ecuación instrumental del problema
Fuente: *BWGHT.RAW (Wooldridge, 2010)*

En el Cuadro 2 se puede notar que el instrumento no es relevante ni en un sentido estadístico ni en un sentido empírico. De hecho, el signo es el opuesto al que se esperaba y ello puede deberse, como se mencionó, a factores regionales que inciden sobre la salud de una madre y por los que no se tomaron controles. Esto, como se evidenció en los resultados del punto *c*), tiene dos efectos sobre la varianza asintótica del estimador **2SLS** presentada en (2.10):

(I) Teniendo en cuenta la ecuación instrumental del problema (2.9), la irrelevancia del instrumento implica que la variabilidad exógena de $packs$, SSE_{packs}, sea menor y, en contrapartida, la pérdida de información relevante de 2SLS (que utilizará SSE_{packs} como $\widehat{SST}_{packs}$) sea mayor. Habrá una fuerte componente exógena de $packs$ no explicada por $cigprice$ que se estará decartando y su efecto sobre (2.10) resulta evidente.

$$packs = \delta_0 + \delta_1 male + \delta_2 parity + \delta_3 log(faminc) + \theta_1 cigprice + r \tag{2.9}$$

(II) La existencia de poca variabilidad en $packs$ adicional a la explicada por el conjunto de variables exógenas del modelo genera una fuerte colinealidad entre las variables, haciendo que $\hat{R}^2_{packs}$, que surge de la proyección lineal de $packs$ sobre el resto de los regresores en (2.8), tienda a 1.

$$Avar(\hat{\beta}^{2SLS}_{packs}) = \frac{\hat{\sigma}^2}{\widehat{SST}_{packs} * (1 - \hat{R}^2_{packs})} \tag{2.10}$$

Este ejercicio nos permite concluir que, a menudo, debemos elegir entre un estimador posiblemente inconsistente, con una varianza menor; y un estimador consistente, pero muy impreciso, de cuyas estimaciones no puede concluirse nada interesante. La utilización de $cigprice$ para instrumentar variabilidad exógena en $packs$ ha demostrado ser una estrategia pobre de identificación para este caso.

"*... it can be very difficult to find a good instrumental variable for an endogenous explanatory variable because the variable must satisfy two different, often conflicting, criteria*" (Wooldridge, 2010, p. 94).

Ejercicio 5.5

La búsqueda de un instrumento para mitigar la inconsistencia del modelo, puede ser problemática. Para el caso de variables omitidas, la estrategia consiste en dejarlas en el error del modelo e instrumentar variabilidad exógena en los regresores endógenos. Para ello, el instrumento debe ser: (*i*) redundante en la ecuación estructural, (*ii*) no debe estar correlacionado con la heterogeneidad inobservable y (*iii*) debe estar parcialmente correlacionado con la variable a instrumentar.

$$Plim\,\hat{\boldsymbol{\beta}}^{IV} - \boldsymbol{\beta} = \left[E(\boldsymbol{x}'\boldsymbol{z}_2)\,E(\boldsymbol{z}_2{}'\boldsymbol{z}_2)^{-1}E(\boldsymbol{z}_2{}'\boldsymbol{x})\right]^{-1} * E(\boldsymbol{x}'\boldsymbol{z}_2)\,E(\boldsymbol{z}_2{}'\boldsymbol{z}_2)^{-1}E(\boldsymbol{z}_2{}'\boldsymbol{u}) \tag{2.11}$$

En donde, siendo $u = \gamma q + a_1$, el estimador IV es consistente en tanto $E(\boldsymbol{z}_2{}'q) = 0$ y $E(\boldsymbol{z}_2{}'a_1) = 0$ (es decir (*i*) y (*ii*)). Adicionalmente, asumiendo que no hay dependencias lineales, $E(\boldsymbol{z}_2{}'\boldsymbol{x})$ es no singular $\Leftrightarrow \boldsymbol{\theta}_{\boldsymbol{z_2}} \neq 0$ (es decir (*iii*)). Como es sabido, no pueden testearse empíricamente debido a la inobservabilidad del error.

$$y_1 = \boldsymbol{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \boldsymbol{z}_2\boldsymbol{\psi}_1 + \gamma q + a_1 \tag{2.12}$$

En (2.12), y_2 es endógena y se utiliza al vector $\boldsymbol{z}_2$ como instrumento. Según la consigna, este último verifica (*i*) $Cov(\boldsymbol{z}_2, a_1) = 0$ y (*iii*) $\theta_{z_2} \neq 0$ pero no se sabe si (*ii*) está parcialmente incorrelacionado con la heterogeneidad inobservable. Si se optara por testear $H_0 : \boldsymbol{\psi}_1 = 0$ en (2.13), el procedimiento carecería de validez.

$$y_1 = \boldsymbol{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \boldsymbol{z}_2\boldsymbol{\psi}_1 + u_1 \tag{2.13}$$

En principio, bajo H_0, $\boldsymbol{z}_1$ y $\boldsymbol{z}_2$ estarían incorrelacionados con u_1, pero la inconsistencia de los estimadores provendría de que y_2 es endógena ($Cov(y_2, q) \neq 0$) y, por tanto, el estadístico de prueba T para testear la exclusión de $\boldsymbol{z}_2$ no es válido.

Si y_2 no fuera endógena, entonces no tendría objeto la introducción de z_2 en el modelo. Si se omitiera a y_2, no se estaría probando la irredundacia de z_2, de hecho, z_2 sería endógena al estar correlacionada con y_2. Una alternativa válida, es la propuesta por Sargan (1958) y Sargan (1988) en donde, si el modelo está sobre-identificado, se puede probar la relevancia de los instrumentos para explicar los residuales de **2SLS**. Si los instrumentos son exógenos en el modelo, los residuales no deben estar correlacionados con los instrumentos.

Ejercicios de Cameron and Trivedi (2005)

Ejercicio 4.7

a)

Proposición 1. *Sean,* $Y = \beta X + u,\ \ X = \lambda u + \varepsilon,\ \ Z = \gamma\varepsilon + v$ *y* $\ u, \varepsilon, v \sim\ iid\ \mathcal{N}[0, \sigma_j^2]$, *entonces:*

$$Plim\,\hat{\beta}^{OLS} - \beta = \ \left[\lambda^2\sigma_u^2 + \sigma_\varepsilon^2\right]^{-1} * \left[\lambda\sigma_u^2\right] \tag{2.14}$$

Demostración. En principio, se deriva el estimador $\hat{\beta}^{OLS}$ para luego, evaluar su inconsistencia. Del modelo planteado para Y se obtiene:

$$\begin{aligned} X'Y &= \beta X'X + X'u \\ X'Y &= \beta X'X + X'u \\ (X'X)^{-1} * (X'Y) &= \beta + (X'X)^{-1} * (X'u) \end{aligned}$$

De esta última, se toma el operador esperanza matemática y se tiene:

$$\beta = E(X'X)^{-1} * E(X'Y)$$

Provisto de que no hay dependencias lineales, el principio de analogía y, por simplicidad, de que el método generalizado de momentos genera el mismo estimador que **OLS**, se tiene que:

$$\hat{\beta}^{OLS} = \left[\frac{\sum_{i=1}^{N} X_i{}'X_i}{N}\right]^{-1} * \frac{\sum_{i=1}^{N} X_i{}'Y_i}{N}$$

$$\hat{\beta}^{OLS} = \beta + \left[\frac{\sum_{i=1}^{N} X_i{}'X_i}{N}\right]^{-1} * \frac{\sum_{i=1}^{N} X_i{}'u_i}{N}$$

Utilizando el hecho de que $X = \lambda u + \varepsilon$, se obtiene:

$$\hat{\beta}^{OLS} = \beta + \left[\frac{\sum_{i=1}^{N} (\lambda u_i' + \varepsilon_i')(\lambda u_i + \varepsilon_i)}{N}\right]^{-1} * \frac{\sum_{i=1}^{N} (\lambda u_i' + \varepsilon_i')u_i}{N}$$

$$\hat{\beta}^{OLS} = \beta + \left[\frac{\sum_{i=1}^{N} (\lambda^2 u_i^2 + \lambda u_i'\varepsilon_i + \lambda\varepsilon_i' u_i + \varepsilon_i^2)}{N}\right]^{-1} * \frac{\sum_{i=1}^{N} (\lambda u_i^2 + \varepsilon_i' u_i)}{N}$$

Tomando el límite probabilístico en esta última ecuación y provisto tanto de la Ley de Débil de los Grandes Números como de la tesis del Teorema de Slutsky, se obtiene:

$$Plim\, \hat{\beta}^{OLS} - \beta = \left[\lambda^2 E(u^2) + \lambda E(u'\varepsilon) + \lambda E(\varepsilon' u) + E(\varepsilon^2)\right]^{-1} * \left[\lambda E(u^2) + E(\varepsilon' u)\right]$$

$$Plim\, \hat{\beta}^{OLS} - \beta = \left[\lambda^2 \sigma_u^2 + \sigma_\varepsilon^2\right]^{-1} * \left[\lambda \sigma_u^2\right] \qquad \square$$

b)

Proposición 2. *Sean,* $Y = \beta X + u,\ X = \lambda u + \varepsilon,\ Z = \gamma\varepsilon + v$ ***y*** $u, \varepsilon, v \sim iid\, \mathcal{N}[0, \sigma_j^2]$, *entonces:*

$$\rho_{X,Z}^2 = \frac{(\gamma\sigma_\varepsilon^2)^2}{(\lambda^2\sigma_u^2 + \sigma_\varepsilon^2)(\gamma^2\sigma_\varepsilon^2 + \sigma_v^2)} \tag{2.15}$$

Demostración. Siendo $\rho_{X,Z}$ el coeficiente de correlación de Pearson, el mismo está dado por:

$$\rho_{X,Z} = \frac{Cov(X,Z)}{\sqrt{V(X) * V(Z)}} \tag{2.16}$$

Calculando cada componente en (2.16):

$$\begin{aligned} Cov(X,Z) &= E[X'Z] - E[X]E[Z] \\ &= E[(\lambda u' + \varepsilon')(\gamma\varepsilon + v)] \\ &= \gamma E[\varepsilon^2] \\ &= \gamma\sigma_\varepsilon^2 \end{aligned} \qquad \begin{aligned} V(X) &= E[X^2] - E^2[X] \\ &= E[(\lambda u + \varepsilon)^2] \\ &= \lambda^2 E[u^2] + E[\varepsilon^2] \\ &= \lambda^2\sigma_u^2 + \sigma_\varepsilon^2 \end{aligned} \qquad \begin{aligned} V(Z) &= E[Z^2] - E^2[Z] \\ &= E[(\gamma\varepsilon + v)^2] \\ &= \gamma^2 E[\varepsilon^2] + E[v^2] \\ &= \gamma^2\sigma_\varepsilon^2 + \sigma_v^2 \end{aligned}$$

Reemplazando los momentos obtenidos en (2.16), $\rho_{X,Z}^2$ está dado por:

$$\rho_{X,Z}^2 = \frac{(\gamma\sigma_\varepsilon^2)^2}{(\lambda^2\sigma_u^2 + \sigma_\varepsilon^2)(\gamma^2\sigma_\varepsilon^2 + \sigma_v^2)} \qquad \square$$

c)

Proposición 3. *Sean,* $Y = \beta X + u,\ X = \lambda u + \varepsilon,\ Z = \gamma\varepsilon + v$ *y* $u, \varepsilon, v \sim iid\, \mathcal{N}[0, \sigma_j^2]$*, entonces el estimador de variables instrumentales está dado por* (2.17) *y en donde, por ejemplo,* $m_{zu} = \sum_i z_i * u_i$*.*

$$\hat{\beta}^{IV} = \frac{m_{zy}}{m_{zx}} = \beta + \frac{m_{zu}}{\lambda m_{zu} + m_{z\varepsilon}} \tag{2.17}$$

Demostración. Para deducir el estimador $\hat{\beta}^{IV}$ partimos del modelo planteado para Y:

$$Z'Y = \beta Z'X + Z'u$$
$$Z'Y = \beta Z'X + Z'u$$
$$(Z'X)^{-1} * (Z'Y) = \beta + (Z'X)^{-1} * (Z'u)$$

De esta última, dado que se satisfacen las condiciones para que Z sea un instrumento válido y relevante (es decir $(Z'X)$ es no singular), se toma el operador esperanza matemática y se tiene:

$$\beta = E(Z'X)^{-1} * E(Z'Y)$$

Provisto de que no hay dependencias lineales y del principio de analogía, se tiene que:

$$\hat{\beta}^{IV} = \left[\frac{\sum_{i=1}^{N} Z_i{}'X_i}{N}\right]^{-1} * \frac{\sum_{i=1}^{N} Z_i{}'Y_i}{N} = \frac{m_{zy}}{m_{zx}} \tag{2.18}$$

Que es una de las expresiones que se querían demostrar. Reemplazando $Y = \beta X + u$ en (2.18) se obtiene:

$$\hat{\beta}^{IV} = \left[\frac{\sum_{i=1}^{N} Z_i{}'X_i}{N}\right]^{-1} * \frac{\beta\left(\sum_{i=1}^{N} Z_i{}'X_i\right) + \sum_{i=1}^{N} Z_i{}'u_i}{N}$$

Utilizando el hecho de que $X_i = \lambda u_i + \varepsilon_i$, se obtiene:

$$\hat{\beta}^{IV} = \left[\frac{\lambda\left(\sum_{i=1}^{N} Z_i{}'u_i\right) + \sum_{i=1}^{N} Z_i{}'\varepsilon_i}{N}\right]^{-1} * \frac{\beta\left[\lambda\left(\sum_{i=1}^{N} Z_i{}'u_i\right) + \sum_{i=1}^{N} Z_i{}'\varepsilon_i\right] + \sum_{i=1}^{N} Z_i{}'u_i}{N}$$

$$\hat{\beta}^{IV} = \beta + \left[\frac{\lambda\left(\sum_{i=1}^{N} Z_i{}'u_i\right) + \sum_{i=1}^{N} Z_i{}'\varepsilon_i}{N}\right]^{-1} * \frac{\sum_{i=1}^{N} Z_i{}'u_i}{N} = \beta + \frac{m_{zu}}{\lambda m_{zu} + m_{z\varepsilon}} \tag{2.19}$$

Las expresiones (2.18) y (2.19) para $\hat{\beta}^{IV}$ se expresan en función de m_{zy}, $m_{z\varepsilon}$ y m_{zx}. □

d) Teniendo en cuenta las hipótesis planteadas en Nelson and Startz (1988):

Proposición 4. *Sean,* $Y = \beta X + u,\ X = \lambda u + \varepsilon,\ Z = \gamma\varepsilon + v$ *y* $u, \varepsilon, v \sim iid\, \mathcal{N}[0, \sigma_j^2]$*. Considerando* ε *y* v *fijas cuando se está muestreando* u *y, adicionalmente* $m_{z\varepsilon} = \gamma\sigma_\varepsilon^2$*, la inconsistencia del estimador converge a* $1/\lambda$ *conforme* $\gamma \to 0$ *o, según* (2.15)*,* $\rho_{X,Z}^2 \to 0$*.*

$$\left(\hat{\beta}^{IV} - \beta\right) \to \frac{1}{\lambda} \quad conforme \quad \gamma \ \ o \ \ \rho_{X,Z} \ \to 0 \tag{2.20}$$

Demostración. De la expresión (2.19), si se consideran fijos ε y v cuando se está muestreando u, resulta que $m_{z\varepsilon}$ es no-estocástico y la única fuente de variabilidad en m_{zu} es u. El valor de $m_{z\varepsilon}$ dependerá de las realizaciones de ε y v, pero estará en mayor o menor medida en torno a $\gamma\sigma_\varepsilon^2$ debido al error de muestreo.

$$\hat{\beta}^{IV} - \beta = \frac{\sum_{i=1}^{N} Z_i{}' u_i}{\lambda \left(\sum_{i=1}^{N} Z_i{}' u_i \right) + \gamma\sigma_\varepsilon^2}$$

Sacando factor común m_{zu}, y reemplazando Z se tiene:

$$\hat{\beta}^{IV} - \beta = \frac{1}{\lambda + \dfrac{\gamma\sigma_\varepsilon^2}{m_{zu}}} \tag{2.21}$$

De esta forma, el tamaño de la inconsistencia de variables instrumentales depende de la realización de m_{zu} cuyo valor esperado es 0. Sin embargo, si su valor es relativamente grande en relación a $\gamma\sigma_\varepsilon^2$, lo cual ocurriría en la medida que $\gamma \to 0$, la inconsistencia converge a $1/\lambda$. □

e)

Proposición 5. *Considerando las hipótesis iniciales y los resultados alcanzados en d), la inconsistencia del estimador tiende a infinito conforme:*

$$\left(\hat{\beta}^{IV} - \beta \right) \to \infty \quad conforme \quad m_{zu} \to \frac{-\gamma\sigma_\varepsilon^2}{\lambda} \tag{2.22}$$

Demostración. De la expresión (2.21), tomando el límite con $m_{zu} \to (-\gamma\sigma_\varepsilon^2)/\lambda$, se obtiene:

$$\hat{\beta}^{IV} - \beta = \lim_{m_{zu}} \frac{1}{\lambda - \dfrac{\lambda\gamma\sigma_\varepsilon^2}{\gamma\sigma_\varepsilon^2}} = \infty$$ □

f) En la práctica, generalmente, los estimadores son inconsistentes. Sin embargo, variables instrumentales nos ofrece una estrategia para atenuar la inconsistencia del estimador en tanto se satisfagan las restricciones de identificación del problema en mayor o menor medida. Se trata de un balance difícil entre **OLS** y **IV** en la medida en que la validez del instrumento no puede probarse y, además, el exacerbamiento de la inconsistencia puede ser aún mayor si el instrumento es invalido o débil.

Los ejercicios presentados en *d*) y *e*) dan cuenta de ello. En la medida en que $\rho_{X,Z}^2 \to 0$, ya sea por $\gamma \to 0$ o por $\sigma_\varepsilon^2 \to 0$, el error del estimador es cada vez mayor hasta converger en el límite a $1/\lambda$. Por otro lado, si el instrumento no fuera válido $m_{zu} \neq 0$, la inconsistencia podría, o bien tender a $1/\lambda$ si $m_{zu} \to \infty$, o bien tender a infinito si $m_{zu} \to (-\gamma\sigma_\varepsilon^2)/\lambda$ agravando aún más el problema. En tales casos, puede ser preferible buscar otros instrumento o seguir con **OLS**.

Ejercicio 8.3

a)

Proposición 6. *Sea $\hat{\theta}$ un estimador asintóticamente eficiente bajo H_0 pero que pierde consistencia bajo H_1 y sea $\tilde{\theta}$ otro estimador que es asintóticamente menos eficiente que el primero bajo H_0 pero permanece consistente bajo H_1. Demostrar que el estimador* (2.23) *tiene por varianza asintótica a* (2.24).[1]

$$\bar{\theta} = \hat{\theta} + [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [\tilde{\theta} - \hat{\theta}] \tag{2.23}$$

$$V(\bar{\theta}) = V_{11} - [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [V_{11} - V_{12}]' \tag{2.24}$$

Siendo $V(\hat{\theta}) = V_{11}$, $V(\tilde{\theta}) = V_{22}$ y $Cov(\hat{\theta}, \tilde{\theta}) = V_{12}$.

Demostración. Tomando la varianza de (2.23), se tiene:

$$V(\bar{\theta}) = V\left[\hat{\theta} + [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [\tilde{\theta} - \hat{\theta}]\right]$$

Dado que no sería correcto suponer independencia entre $\hat{\theta}$ y $\tilde{\theta} - \hat{\theta}$ se tiene:

$$\begin{aligned} V(\bar{\theta}) = V[\hat{\theta}] &+ V\left[[V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [\tilde{\theta} - \hat{\theta}]\right] \\ &+ 2 * [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * Cov\left(\hat{\theta}; [\tilde{\theta} - \hat{\theta}]\right) \end{aligned}$$

Siendo $[V_{11} - V_{12}] * [\cdots]^{-1}$ matrices no estocásticas, la única variable aleatoria es $\tilde{\theta} - \hat{\theta}$.

$$\begin{aligned} V(\bar{\theta}) = V[\hat{\theta}] &\\ &+ [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * V[\tilde{\theta} - \hat{\theta}] * \left([V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1}\right)' \\ &+ 2 * [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * Cov\left(\hat{\theta}; [\tilde{\theta} - \hat{\theta}]\right) \end{aligned} \tag{2.25}$$

Por motivos expositivos, se procede a desarrollar $Cov\left(\hat{\theta}; [\tilde{\theta} - \hat{\theta}]\right)$ por separado.

$$\begin{aligned} Cov\left(\hat{\theta}; [\tilde{\theta} - \hat{\theta}]\right) &= E[\hat{\theta}(\tilde{\theta} - \hat{\theta})] - E[\hat{\theta}] * E[\tilde{\theta} - \hat{\theta}] \\ &= E[\hat{\theta}\tilde{\theta}] - E[\hat{\theta}] * E[\tilde{\theta}] + E^2[\hat{\theta}] - E[\hat{\theta}^2] \\ &= Cov(\tilde{\theta}; \hat{\theta}) - V(\hat{\theta}) \end{aligned}$$

Que convenientemente hemos nombrado:

$$Cov\left(\hat{\theta}; [\tilde{\theta} - \hat{\theta}]\right) = V_{12} - V_{11} \tag{2.26}$$

Reemplazando (2.26) en (2.25) y sacando factor común (-1) se obtiene:

$$\begin{aligned} V(\bar{\theta}) = V[\hat{\theta}] &\\ &+ [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * V[\tilde{\theta} - \hat{\theta}] * \left([V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1}\right)' \\ &- 2 * [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [V_{11} - V_{12}]' \end{aligned}$$

[1] Este ejercicio está inspirado en Amemiya (1985, p. 146) para evaluar las dinámicas del test de Hausman. Dado que se han encontrado varios errores de tipeo en la consigna 8.3 de Cameron and Trivedi (2005, p. 293), se ha recurrido al libro original en pos de corregir los errores y hacer correctamente el ejercicio.

En donde $[V_{11} + V_{22} - 2V_{12}]^{-1} * V[\tilde{\theta} - \hat{\theta}] = I$ y se desarrolla la traspuesta del producto de matrices, resulta:

$$\begin{aligned} V(\bar{\theta}) = V_{11} \\ &+ [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [V_{11} - V_{12}]' \\ &- 2 * [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [V_{11} - V_{12}]' \end{aligned}$$

Que no es otra cosa que:

$$V(\bar{\theta}) = V_{11} - [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [V_{11} - V_{12}]' \qquad \square$$

b)

Proposición 7. *Considerando los resultados alcanzados en a),* $V(\bar{\theta})$ *es menor* $V(\hat{\theta})$ *en términos matriciales a menos que* $Cov(\hat{\theta}; \tilde{\theta}) = V(\hat{\theta}) = V_{11}$.

Demostración.

$$V(\hat{\theta}) = V_{11} \tag{2.27}$$

$$V(\bar{\theta}) = V_{11} - [V_{11} - V_{12}] * [V_{11} + V_{22} - 2V_{12}]^{-1} * [V_{11} - V_{12}]' \tag{2.28}$$

Dado que toda matríz de covarianzas es semi-definida positiva y que aquí estamos admitiendo su inversa, entonces $[V_{11} + V_{22} - 2V_{12}]$ es definida positiva y, por tanto, su inversa también los es. Por otro lado, asumiendo que $[V_{11} - V_{12}]$ es no nulo, el segundo sumando en (2.28) es siempre positivo. Esto quiere decir que, asintóticamente, la varianza de $\bar{\theta}$ menor a la de $V(\hat{\theta})$. La única forma de que esto no suceda es si $[V_{11} - V_{12}]$ es nulo o, en otros términos, cuando $Cov(\hat{\theta}; \tilde{\theta}) = V_{11}$. $\square$

c) Por más que la matríz $[V_{11} + V_{22} - 2V_{12}]$ sea definida positiva, si el estimador $\hat{\theta}$ resulta ser eficiente, entonces $Cov(\hat{\theta}; \tilde{\theta}) = V_{11}$ y, según la expresión (2.28), debe verificarse asintóticamente:

$$V(\bar{\theta}) = V(\hat{\theta}) \tag{2.29}$$

3. Ejercicio 3 - Curva Salarial Estática

En este apartado se reproducen los resultados empíricos para la economía argentina en el período 2003-2014 contenidos en Card (1995), quien hace una revisión del trabajo de Blanchflower and Oswald (1989), y recopila los contenidos en las Tablas I, II, III y IV.

En cuanto al caso argentino, al hacer uso de una base de datos de **pool de corte transversal**, surge un inconveniente para reproducir la Tabla 1 de Card (1995). Esta, recopila resultados de cuatro papers diferentes que utilizan observaciones repetidas del salario para los mismos individuos, es decir la estructura de los datos es **datos de panel**. Por lo tanto, ya que la base de datos disponible no permite reproducir esos resultados, procedemos a especificar 4 modelos diferentes para evaluar como varía la elasticidad de los salarios reales respecto al desempleo en cada uno de ellos.

```
1) regress lw educ lu exper expersq fem single reg0-reg4 year
2) regress lw educ lu exper expersq fem single reg0-reg4 year2004-year2014
3) regress lw educ lu exper expersq fem single reg0-reg4 year educyear
4) regress lw educ lu exper expersq fem single reg0-reg4 year2004-year2014
educyear2004-educyear2014
```

La diferencia entre la primera y segunda especificación es el uso de una variable nominal (*year*) y de variables binarias o dummies (*year2004-year2014*). Por otro lado, la tercera y cuarta especificación incorporan términos de interacción entre las variables *year* y *educ*, pero difieren que en una *year* es una variable nominal mientras que en otra *year2004-year2014* son variables dummies.

La razón para incorporar variables dummies o términos de interacción es para introducir una no linearidad. En el caso de las variables nominales, se presenta una restricción de equidistancia entre las múltiples categorías, es decir que hay linearidades en los efectos, lo cual es un supuesto poco realista. Por ello, es posible descomponer una variable nominal en variables binarias y así eliminar el supuesto de equidistancia, introduciendo hasta J-1 variables binarias para evitar la trampa de la variable dummy.

Si bien la metodología utilizada para confeccionar esta tabla difiere de la del artículo, podemos ver que para el caso argentino se verifica una elasticidad negativa salario-desempleo al igual que en trabajo mencionado. Asimismo, los coeficientes son significativos, con un *p-valor¡0.001*, lo que determina la robustez de las estimaciones. Podemos destacar a su vez la similitud de los coeficientes, en especial para los casos de las especificaciones (2) y (4) del Cuadro 3, donde un aumento del 1 % del desempleo trae aparejada una caída porcentual del salario del -0.094 y -0.0936, respectivamente, en relación con los valores entre -0.081 y -0.089 en las estimaciones para Estados Unidos.

	[1]	**[2]**	**[3]**	**[4]**
	Log(w)	Log(w)	Log(w)	Log(w)
Log(U)	-0.0626***	-0.0940***	-0.0617***	-0.0936***
	(0.00322)	(0.00335)	(0.00322)	(0.00335)

Cuadro 3: Tabla 1, Elasticidad salarios reales-desempleo agregado para Argentina.
Fuente: *Unidad individual 2003-2014.*

Para realizar la estimación de la Tabla 2, Card describe dos metodologías diferentes que adoptan en Blanchflower and Oswald (1989), que utilizan repeticiones de cortes transversales en lugar de datos de panel, como las estimaciones resumidas en la Tabla 1.

En la **primera metodología**, la estimación se realiza usando un pool de micro datos de corte transversal. La estimación de las elasticidades *salario-desempleo* se realizan incluyendo controles por el salario real observado en el mercado laboral local del individuo, la tasa de

desempleo en ese mercado y a la vez se incluyen otras características de los individuos, tales como el género, edad, educación. Para la estimación en el caso argentino, se utilizan dichos controles a la vez que se añaden características como la experiencia, el estado civil y la región.

Los autores aproximan la curva salarial mediante una función log-linear, similar a la que empleamos para las estimaciones del caso argentino. La regresión utilizada para la replicar la estimación con micro datos de la Tabla 2 para el caso argentino es la siguiente:

```
regress lw lu educ exper expersq age fem single reg0-reg4  year2004-year2014
```

Se utilizan variables dummy tanto para las regiones como para los años presentes en la muestra. Si las variables dummy de mercado fueran excluidas, la asunción implícita sería que los salarios responden a los componentes transitorios y permanentes del desempleo local con la misma elasticidad.

Como la primera especificación permitía que individuos en el mismo mercado laboral compartiesen componentes comunes de varianza no enteramente atribuibles a sus características o la tasa de desempleo local, los errores de la estimación estaban correlacionados positivamente entre individuos del mismo mercado laboral. Blanchflower y Oswald son conscientes de este problema y utilizan una **segunda metodología**. Esta consiste en un procedimiento de agregación simple, en la cual se toman los promedios de los individuos en un mismo mercado y período, correspondiente a la columna "cell means" de la Tabla 2. Mediante esta metodología, se asume que no hay correlación en los determinantes inobservables de los salarios entre mercados y los residuos de la estimación no estarán correlacionados entre observaciones.

Los comandos utilizados para replicar la estimación correspondiente a "cell means" de la Tabla 2 para el caso argentino son los siguientes:

```
collapse (mean) lw educ age exper expersq lu, by(region year),
```

mediante la cual realizamos el promedio de las observaciones de los distintos individuos en la misma región y por período. Luego realizamos la regresión

```
regress lw lu exper expersq age educ i.region i.year
```

A diferencia de Card, que utiliza dos estimaciones agregadas con tasa de desempleo regional e industrial, para el caso argentino utilizamos únicamente la tasa de desempleo regional, ya que es la única con la que contamos.

Los resultados arrojados con los datos de Argentina para el período 2003-2014, son muy similares a las obtenidas por Blanchflower y Oswald, tanto para la estimación con micro datos como para aquella obtenida mediante el promedio de las observaciones individuales.

En ambos casos, los coeficientes para la muestra argentina confirman la elasticidad negativa entre salarios y desempleo (-0.094 y -0.0756) y muestran una notable significatividad con p-valores menores a 0.001 y 0.01 respectivamente. Se cumple a su vez, en concordancia con el caso del paper, que en la estimación mediante agregación los errores estándar sean mayores y las elasticidades sean menores.

	Micro Estimates	**Cell Means**
Log(U)	-0.0940***	-0.0756**
	(0.00335)	(0.0265)

Cuadro 4: Tabla 2, Elasticidad de la curva de salario estimada para Argentina

Fuente: *Unidad individual 2003-2014.*

Al confeccionar la Tabla 3, Card utiliza dos medidas diferentes: un promedio simple de las variables para los trabajadores de una región (Actual) y una regresión ajustada donde controla por las características observables de los trabajadores en cada región (Ajustado). Este análisis

lo lleva a cabo para calcular las elasticidades del salario por hora, las horas anuales trabajadas y los ingresos anuales con respecto a la tasa de desempleo. La estimación la realiza para distintos subgrupos, lo que permite calcular diferencias entre grupos para las elasticidades. Sin embargo, el análisis realizado para el caso argentino estima la elasticidad del salario con respecto al desempleo, por ser los datos disponibles en la muestra.

	Actual		**Ajustada**	
Total	-0.0756	(0.0265)	-0.0940	(0.00335)
Por género:				
Mujer	-0.0209	(0.0294)	-0.0951	(0.00519)
Hombre	-0.0982	(0.0274)	-0.0932	(0.00437)
Por estado civil:				
Soltero	-0.0554	(0.0330)	-0.0854	(0.00613)
Casado	-0.0880	(0.0253)	-0.0968	(0.00399)
Por edad:				
Edad 16-29	-0.0954	(0.0382)	-0.0847	(0.00618)
Edad 30-44	-0.0941	(0.0342)	-0.0843	(0.00517)
Edad $>$45	-0.0418	(0.0326)	-0.111	(0.00606)
Por educación:				
Educ $<$12	-0.0848	(0.0414)	-0.0961	(0.00513)
Educ 12-15	-0.0663	(0.0233)	-0.0829	(0.00539)
Educ +16	-0.0422	(0.0380)	-0.107	(0.00752)

Cuadro 5: Tabla 3, Elasticidad salarios-desempleo por tipo de trabajador.
Fuente: *Unidad individual 2003-2014.*

La regresión para la estimación ajustada es la misma que la utilizada para la estimación con micro datos de la Tabla 2, agregando condicionantes para los subgrupos en cuestión, por ejemplo, if fem==1 en el caso de mujeres, if single==0 para casados o if educ¿15 para aquellos individuos con 16 años o más de educación. En cambio, la regresión utilizada para la agrupación simple se asemeja a aquella utilizada en la Tabla 2 para el campo “cell means”, añadiendo condicionantes según el subgrupo analizado.

Por último, para evaluar si el cambio del desempleo actual tiene el mismo efecto en los salarios de diferentes tipos de trabajadores, o diferentes industrias o sectores, Card presenta un resumen de estudios empíricos, llevados a cabo por Blanchflower y Oswald para distintos países, en la Tabla 4. Para evaluar el caso argentino, analizamos el efecto del desempleo en los salarios de diferentes grupos de trabajadores (clasificados por género, estado civil, edad, años de educación) o regiones (NOA, NEA Cuyo, Pampeana y Patagónica), y como la muestra es de un mismo país, la dividimos por períodos (2003-2005, 2006-2008, 2009-2011, 2012-2014). A diferencia de la tabla 4 del trabajo, no se dispone de información acerca de la afiliación sindical ni pertenencia al sector laboral privado o público.

Las conclusiones que se desprenden de las Tablas 3 y 4 para la economía argentina en cuanto a género son contrarias al trabajo de Card. En este, los salarios de los hombres son más sensibles a las tasas de desempleo local que el salario de las mujeres. En cambio, en Argentina, en el período inicial de la muestra (2003-2005), la elasticidad es mayor en el caso de las mujeres (-0.11 contra -0.0658), luego se equipara a la elasticidad de los hombres para los períodos 2006-2008 y 2009-2011, es decir que se reduce la brecha de género, para finalmente situarse debajo de la elasticidad de los hombres para el período 2012-2014 (-0.0483 y -0.0599 respectivamente).

Con respecto al estado civil, no se observa una diferencia significativa entre las elasticidades salario-desempleo de solteros y casados en los primeros dos períodos. Sin embargo, hacia el final de la muestra, la elasticidad de los soleteros comienza a descender más deprisa

	2003-2005		2006-2008		2009-2011		2012-2014	
Todo	-0.0849	(0.00877)	-0.125	(0.00691)	-0.100	(0.00615)	-0.0553	(0.00620)
Mujer	-0.110	(0.0131)	-0.120	(0.0107)	-0.105	(0.00963)	-0.0483	(0.00989)
Hombre	-0.0658	(0.0118)	-0.129	(0.00904)	-0.0957	(0.00797)	-0.0599	(0.00792)
Soltero	-0.101	(0.0160)	-0.132	(0.0125)	-0.0845	(0.0114)	-0.0279	(0.0115)
Casado	-0.0783	(0.0105)	-0.121	(0.00828)	-0.105	(0.00732)	-0.0660	(0.00737)
Edad 16-29	-0.0972	(0.0152)	-0.128	(0.0123)	-0.0777	(0.0115)	-0.0181	(0.0122)
Edad 30-44	-0.0487	(0.0136)	-0.120	(0.0109)	-0.0999	(0.00958)	-0.0429	(0.00936)
Edad>45	-0.110	(0.0166)	-0.126	(0.0126)	-0.116	(0.0110)	-0.0911	(0.0110)
Educ<12	-0.0998	(0.0122)	-0.149	(0.0105)	-0.0815	(0.00964)	-0.0480	(0.0101)
Educ 12-15	-0.0704	(0.0153)	-0.1000	(0.0112)	-0.105	(0.00986)	-0.0432	(0.00963)
Educ +16	-0.0599	(0.0218)	-0.124	(0.0159)	-0.125	(0.0134)	-0.0902	(0.0133)
NOA	-0.181	(0.0290)	-0.00481	(0.0183)	-0.0622	(0.0217)	0.0171	(0.0245)
NEA	-0.118	(0.0328)	-0.199	(0.0322)	-0.130	(0.0156)	-0.102	(0.0164)
Cuyo	-0.0919	(0.0150)	-0.0524	(0.0127)	-0.0499	(0.0127)	-0.0697	(0.0125)
Pampeana	-0.0560	(0.0216)	0.00754	(0.0195)	-0.0821	(0.0113)	-0.0415	(0.00900)
Patagonica	0.0285	(0.0174)	-0.252	(0.0117)	-0.158	(0.0145)	0.0290	(0.0275)

Standard errors in parentheses

Cuadro 6: Tabla 4, Elasticidades curva de salario por tipo de trabajador y sector para Argentina.
Fuente: *Unidad individual 2003-2014.*

que la de los casados, alcanzando menos de la mitad del valor de la elasticidad de los casados para el periodo 2012-2014 (-0.0279 contra -0.066). Asimismo, como se observa en la Tabla 3, las elasticidades de los mayores de 45 años (-0,111), como los de aquellos con más de 16 años de educación (-0.107), son sustancialmente menores que las de individuos más jóvenes y con menos años de educación, lo cual concuerda con los hallazgos del trabajo de Card. Por último, en cuanto a las diferentes regiones del país, todas presentan coeficientes negativos de elasticidades salario-desempleo, a excepción de la Patagonia, que presenta elasticidades positivas para los períodos 2003-2005 y 2012-2014 (0.0285 y 0.0290 respectivamente). La región cuyo salario resulta más vulnerable al desempleo a lo largo de los diferentes periodos es el NEA, siendo la región Pampeana la menos vulnerable y la región Patagónica aquella cuyas elasticidades son más volátiles (0.0285, -0.252, -0.158, 0.029 para los cuatro períodos).

4. Ejercicio 4 - Perfil de la Informalidad

El ejercicio realizado busca encontrar patrones en el perfil de informalidad laboral a partir de la EAHU que realiza el INDEC. Se trata de una muestra aleatoria, estratificada y polietápica en el que, al igual que en la EPH, los hogares encuestados siguen un esquema de rotación: de un año al otro existe aproximadamente un 50 % de la muestra común al año anterior. Esto es relevante dado que es muy infrecuente la existencia de **datos de panel** siendo la excepción países desarrolados. En Argentina, con este tipo de encuestas, podrían identificarse los hogares que se repiten en la muestra de un año a otro pero solo sería por un periodo corto de tiempo. Es por ello que, en este trabajo, se utiliza una estructura de datos de tipo **pool de corte transversal**.

Para dar cuenta del perfil de los trabajadores informales, se utilizaran dos modelos para predecir la variable:

$$Y_i = \begin{cases} 1 & tiene\ descuento\ jubilatorio \\ 0 & no\ tiene\ descuento\ jubilatorio \end{cases} \tag{4.1}$$

Los modelos a evaluar son dos: un modelo de probabilidad lineal (**LPM**) y un modelo con variable dependiente limitada **Probit**. El objetivo primero del análisis será encontrar los factores que inciden sobre la probabilidad de tener un trabajo informal y detectar posibles prácticas discriminatorias en el mercado laboral. Para ello se utilizan variables *dummy* como *migrante sudamericano*, *mujer*, entre otras; controlando por otros factores que afectan la probabilidad de ser informal y que podrían estar correlacionadas con estas, por ejemplo, el nivel de estudios alcanzado, la región del hogar, la actividad económica en la que se emplea, la edad, etc.

4.1. Modelo de probabilidad lineal

Para este caso se modela la variable dependiente como una variable aleatoria de Bernoulli en donde su función de probabilidad condicional está dada por (4.2) y, por tanto, estaremos modelando la probabilidad de éxito Y_i condicional en los regresores, según (4.3).

$$Y_i \sim \mathcal{B}ernoulli(P(\boldsymbol{X}))$$

$$P[Y_i = y_i|\boldsymbol{X}] = P^{y_i}(\boldsymbol{X}) * [1 - P(\boldsymbol{X})]^{1-y_i} \tag{4.2}$$

$$E[Y_i|\boldsymbol{X}] = P(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta} \tag{4.3}$$

Las ventajas de este tipo de modelos son varias pero, dependiendo de la pregunta de investigación, puede ser inapropiado. En principio (*i*) **OLS**, bajo las restricciones de identificación habituales, genera estimadores consistentes para $\boldsymbol{\beta}$ en (4.3), (*ii*) la interpretación del cambio en la probabilidad de ser un trabajador informal dado un aumento unitario en alguno de sus regresores es directa (pues, $\partial E[Y_i|\boldsymbol{X}]/\partial x_j = \beta_j$) y (*iii*) si bien el modelo es por construcción heterocedástico, los estadísticos robustos t, LM y F usuales son válidos.

En el Cuadro 7, algo que llama la atención es la significatividad de los casi todos regresores en un sentido estadístico. Ello no debe ser relevante dado que no se computó la matriz de covarianzas con la corrección propuesta por White (1980) y los estadísticos del test no son válidos. Aún así, los estimadores son asintóticamente consistentes y, por ejemplo, puede notarse que existe una penalidad por sexo en el mercado laboral: el hecho de ser mujer, aumentaba la probabilidad de ser un trabajador informal en 8 puntos porcentuales (p.p.) en 2011 –controlando por nivel de estudios, región, edad y actividad económica– y tal penalización no se redujo significativamente en esta ventana temporal. Por otro lado, la penalidad por ser inmigrante sudamericano se ha reducido drásticamente desde 2011 aún cuando su participación en la población activa no se ha reducido.

Variable	O2011	O2012	O2013	O2014
Migrante Sudamericano	0.08**	0.10**	0.06	0.02
= 1 Mujer	0.08***	0.07***	0.09***	0.07***
De 25 a 49 Años	-0.14***	-0.17***	-0.16***	-0.16***
De 50 Años y Más	-0.11***	-0.15***	-0.13***	-0.16***
Primario Completo	-0.08***	-0.08***	-0.01	-0.12***
Secundario Incompleto	-0.09***	-0.10***	-0.04	-0.11***
Secundario Completo	-0.20***	-0.23***	-0.16***	-0.22***
Universitario Incompleto	-0.23***	-0.25***	-0.20***	-0.22***
Universitario Completo	-0.28***	-0.30***	-0.24***	-0.31***
NOA	0.07***	0.06***	0.06***	0.08***
NEA	0.06***	0.08***	0.09***	0.10***
Cuyo	0.05***	0.03**	0.01	0.02
Pampeana	0.02	0.02	0.01	0.03*
Patagónica	-0.09***	-0.09***	-0.12***	-0.10***
Actividades Primarias	0.08***	0.09***	0.10	0.14***
Explotación Minas y Canteras	-0.13***	-0.10***	-0.11***	-0.11***
Industria Manufacturera	-0.00	0.01	0.03	0.04**
Suministro de Electricidad y Gas	-0.14***	-0.15***	-0.12***	-0.09***
Suministro de Agua	-0.02	-0.08*	0.01	0.07
Construcción	0.23***	0.25***	0.31***	0.31***
Comercio	0.06***	0.12***	0.11***	0.12***
Transporte y Almacenamiento	0.11***	0.12***	0.15***	0.08*
[3;6] Meses	-0.16***	-0.18***	-0.10***	-0.17***
[6;12] Meses	-0.26***	-0.24***	-0.25***	-0.28***
[1;5] Años	-0.36***	-0.37***	-0.35***	-0.37***
$>$5 Años	-0.52***	-0.52***	-0.51***	-0.50***
Constant	0.90***	0.95***	0.86***	0.93***
N	11582873	11956623	12115128	11759149
R^2	0.32	0.36	0.35	0.34
			legend: *	p$<$0.001

Cuadro 7: Modelo para Asalariados Informales OLS - Discriminación para 2011-2014
Fuente: *Encuesta Anual de Hogares Urbanos, INDEC.*

Al observar los signos de las variables binarias asociadas a la edad, puede verse que los adultos menores de 25 años son los que presentan una mayor exposición al riesgo de informalidad, siendo la reducción del riesgo cercana a 14 p.p. en 2011 y llegando a 16 p.p. hacia 2014 *vis-à-vis* las demás características de los individuos. Puede observarse que el modelo pronostica bien en la muestra, explicando un 32 % de la variabilidad de Y. Sin embargo, mas que realizar buenos pronósticos, el objetivo es una estimación consistente de efectos específicos asociados a la discriminación en el mercado laboral.

Argumentos en contra de LPM

Una de las críticas más importantes que se le puede hacer al modelo está asociada a que el efecto parcial en el margen de cada regresor sobre la probabilidad de ser informal es constante sin importar el nivel de las variables. Ello podría ser una dificultad en otro tipo de especificaciones para Y, pero aquí se han utilizado regresores binarios que introducen no linealidades en los efectos parciales asociados a la edad, región geográfica, entre otros. La única dificultad asociada a este hecho estriba en que habrá observaciones para las cuales el modelo pronostica probabilidades fuera del intervalo $[0, 1]$, aunque pueden solucionarse con ajustes *ad-hoc* a los pronósticos del modelo.

4.2. Modelo Probit

Para este caso, se utiliza un modelo que restringe la forma en que la variable de respuesta binaria depende de los regresores según (4.4), siendo el modelo **Probit** el caso particular en

que $G(z)$ está dada por (4.5).

$$P[y=1|\boldsymbol{X}] = G(\boldsymbol{X}\boldsymbol{\beta}) \qquad con\ \ 0 < G(z) < 1\ \ \forall z \epsilon \mathbb{R}c \tag{4.4}$$

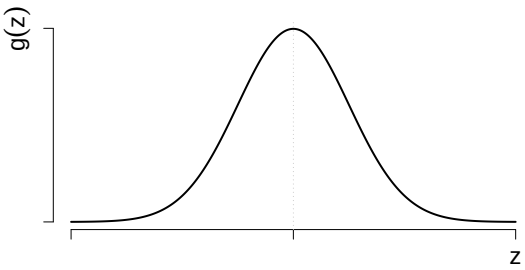
$$G(z) = \int_{-\infty}^{z} \frac{e^{-v^2/2}}{\sqrt{2\pi}} dv \tag{4.5}$$

Las ventajas de este tipo de modelos están en mayor o menor medida relacionadas a cual sea la pregunta de investigación. Siendo el objetivo principal explicar los efectos de $\boldsymbol{X}$ sobre la probabilidad $P[y=1|\boldsymbol{X}]$ las ventajas son: **Probit** realiza pronósticos dentro del intervalo $[0,1]$ y permite la estimación de efectos parciales no lineales sobre la probabilidad de informalidad. Ello quiere decir que, por ejemplo, un hijo adicional en una familia puede tener una incidencia diferente sobre el riesgo de informalidad de los padres dependiendo del nivel de hijos que ya se tengan, nivel de ingresos, edad, etc. El nivel de cada una de las variables tendrá incidencia sobre el efecto parcial en el margen de los demás factores según (4.6), si x_k es una variable contínua, o (4.7) si se trata de una variable binaria, como es el caso.

$$\frac{\partial P[y=1|\boldsymbol{X}]}{\partial x_j} = g(\boldsymbol{X}\boldsymbol{\beta}) * \beta_j \tag{4.6}$$

$$G(\beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + \beta_k) - G(\beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}) \tag{4.7}$$

Adicionalmente, si bien los parámetros del modelo no tienen una interpretación directa como en **LPM**, se puede saber el signo del impacto de cada regresor sobre la $P(\boldsymbol{X})$ tan solo viendo el signo de β_j dado que $g(\boldsymbol{X}\boldsymbol{\beta}) > 0$, siempre. Dado que se trata de un modelo no lineal, la



estimación de los parámetros es por el método de máxima verosimilitud, donde $\hat{\boldsymbol{\beta}}^{Probit}$ verifica (4.8), siendo $\mathcal{L}[\boldsymbol{X};\boldsymbol{\beta}]$ la función de verosimilitud de la muestra.

$$\hat{\boldsymbol{\beta}}^{Probit} = \underset{\boldsymbol{\beta}}{\arg\max}\ \ \mathcal{L}[\boldsymbol{X};\boldsymbol{\beta}] \tag{4.8}$$

Bajo los supuestos del modelo, $\hat{\boldsymbol{\beta}}^{Probit}$ es consistente, asintóticamente normal y con una varianza asintótica que puede usarse para construir estadísticos de prueba t e intervalos de confianza asintóticamente válidos.

En el Cuadro 8 puede notarse que el signo de los efectos es consistente con el **LPM** presentado en el Cuadro 7 y que la *pseudo-R-cuadrada*, definida como $1-(\mathcal{L}_{nr}/\mathcal{L}_0)$, presenta una bondad de ajuste similar en la muestra.

Dado que el efecto parcial de cada variable es no lineal en $\boldsymbol{X}$ debieran elegirse los niveles de interés sobre los cuales estimar el efecto específico de cada una sobre la probabilidad de ser un asalariado informal. Una práctica usual es computar el **efecto parcial en la media** (PEA), que evalúa cada x_j de X en su media muestral, o bien el **efecto parcial promedio** (APE), que promedia el efecto parcial sobre cada individuo en la muestra (realizando, en síntesis, un ejercicio contrafactual para cada individuo). Mediante el comando mkf de Stata, se puede computar el PEA.

Variable	P2011	P2012	P2013	P2014
Migrante Sudamericano	0.26*	0.34**	0.20	0.09
= 1 Mujer	0.27***	0.24***	0.32***	0.23***
De 25 a 49 Ańos	-0.43***	-0.53***	-0.51***	-0.50***
De 50 Ańos y Más	-0.33***	-0.47***	-0.40***	-0.49***
Primario Completo	-0.28***	-0.28***	-0.02	-0.42***
Secundario Incompleto	-0.31***	-0.34***	-0.10	-0.40**
Secundario Completo	-0.67***	-0.78***	-0.52***	-0.75***
Universitario Incompleto	-0.79***	-0.84***	-0.67***	-0.75***
Universitario Completo	-0.97***	-1.11***	-0.89***	-1.17***
NOA	0.23***	0.24***	0.23***	0.30***
NEA	0.21***	0.32***	0.32***	0.37***
Cuyo	0.18***	0.15**	0.03	0.08
Pampeana	0.06	0.08	0.01	0.11*
Patagónica	-0.36***	-0.40***	-0.54***	-0.42***
Actividades Primarias	0.25***	0.30***	0.38*	0.46***
Explotación Minas y Canteras	-1.00***	-0.72***	-1.49***	-0.77***
Industria Manufacturera	0.01	0.08	0.15*	0.16**
Suministro de Electricidad y Gas	-0.77***	-1.21***	-1.07***	-0.67**
Suministro de Agua	-0.08	-0.31	0.02	0.24
Construcción	0.74***	0.85***	1.04***	1.03***
Comercio	0.23***	0.43***	0.40***	0.42***
Transporte y Almacenamiento	0.40***	0.46***	0.55***	0.33**
[3;6] Meses	-0.47***	-0.57***	-0.30***	-0.51***
[6;12] Meses	-0.74***	-0.73***	-0.74***	-0.84***
[1;5] Ańos	-1.02***	-1.10***	-1.00***	-1.08***
$>$5 Ańos	-1.60***	-1.67***	-1.61***	-1.57***
Constant	1.21***	1.39***	1.08***	1.34***
N	11582873	11956623	12115128	11759149
R^2_{pseudo}	0.27	0.30	0.30	0.29
			legend: *	p$<$0.001

Cuadro 8: Modelo para Asalariados Informales PROBIT - Discriminación para 2011-2014

Fuente: *Encuesta Anual de Hogares Urbanos, INDEC.*

y = Pr(Y) (predict)
= .32034958

Variable	dy/dx	Std. Err.	z	P$>\|z\|$	X
Migrante Sudamericano*	.0336406	.03576	0.94	0.347	.034925
= 1 Mujer*	.0828989	.01549	5.35	0.000	.43979
De 25 a 49 Años*	-.1840185	.02142	-8.59	0.000	.634844
De 50 Años y Más*	-.1629407	.02152	-7.57	0.000	.213465
Primario Completo*	-.1404779	.03718	-3.78	0.000	.1815
Secundario Incompleto*	-.1326626	.03849	-3.45	0.001	.175725
Secundario Completo*	-.2382128	.03358	-7.09	0.000	.261096
Universitario Incompleto*	-.2228659	.03078	-7.24	0.000	.120311
Universitario Completo*	-.3311272	.02667	-12.41	0.000	.209418
NOA*	.1138672	.01763	6.46	0.000	.101624
NEA*	.1386814	.02222	6.24	0.000	.067002
Cuyo*	.030853	.02075	1.49	0.137	.06313
Pampeana*	.0389177	.01875	2.08	0.038	.191033
Patagónica*	-.1360004	.01629	-8.35	0.000	.053247
Actividades Primarias*	.1783099	.043	4.15	0.000	.020234
Explotación Minas y Canteras*	-.2123028	.03651	-5.82	0.000	.006709
Industria Manufacturera*	.0599027	.01928	3.11	0.002	.127075
Suministro de Electricidad y Gas*	-.1927371	.04353	-4.43	0.000	.005889
Suministro de Agua*	.0908935	.06138	1.48	0.139	.006114
Construcción*	.3919414	.04065	9.64	0.000	.075988
Comercio*	.1589767	.02342	6.79	0.000	.128553
Transporte y Almacenamiento*	.1232691	.04574	2.69	0.007	.060332
[3;6] Meses*	-.1590465	.02483	-6.40	0.000	.035251
[6;12] Meses*	-.2334668	.02547	-9.17	0.000	.051954
[1;5] Años*	-.3308519	.01782	-18.56	0.000	.293019
$>$5 Años*	-.4918133	.01697	-28.99	0.000	.419675

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Cuadro 9: Marginal effects after dprobit

Fuente: *Encuesta Anual de Hogares Urbanos, INDEC.*

En el Cuadro 9 se obtuvieron los efectos parciales evaluados en la media para el año 2014 y, como puede observarse, las conclusiones no difieren demasiado de las presentadas en el Cuadro 7. Si nos enfocamos en las variables relevantes al problema, la penalidad por sexo en el mercado laboral se estima en torno a 0.02 p.p. para el **LPM** y en 0.03 p.p. para el **Probit**; el perfil jóven de la informalidad laboral sigue siendo notorio con una reducción de 16 y 18 p.p. al pasar al estrato etario de 25 a 49 años para **LPM** y **Probit**, respectivamente. Por último, la discriminación por nacionalidad tiene cierta discrepancia y se presenta en la Figura 1.

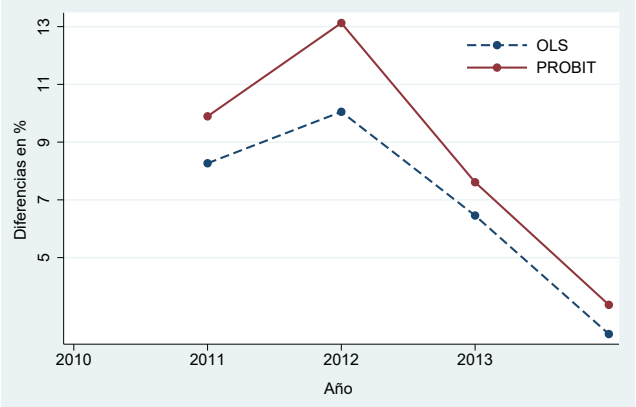


Figura 1: Discriminación entre migrantes y nativos
Fuente: *Encuesta Anual de Hogares Urbanos, INDEC.*

4.3. Teoría y evidencia: comentarios finales

En términos comparativos, ningún modelo es mejor que otro, sino que depende de la pregunta de investigación que se ha planteado. Si bien el modelo **Probit** es extremadamente útil para la estimación de efectos parciales no lineales, esta característica no puede ser explotada con la especificación planteada en este ejercicio, dado que todas las variables son binarias y el **LPM** puede captar perfectamente los mismos efectos. De hecho, ha de destacarse, que si de la estimación del APE se trata, el **LPM** puede estimarlo consistentemente y será un buen predictor para desviaciones moderadas de $\boldsymbol{X}$ de la normalidad (Wooldridge, 2010, p. 579).

Puede que no sea un buen predictor para individuos situados en valores extremos, pero podrá predecir la media poblacional consistentemente bajo los supuestos convencionales de OLS. Por el contrario, las inconsistencias de **Probit** pueden provenir incluso de variables omitidas tanto correlacionadas con los regresores como independientes de ellos (este último problema conocido como *neglected heterogeneity*. Sin embargo, nos permite estimar consistentemente el APE.).

En términos de la evidencia encontrada, si bien discriminación por nacionalidad (sudamericana de países limítrofes) ha mostrado una fuerte reducción a lo largo de los años, es destacable la persistencia de la penalidad por sexo en el mercado laboral y la fuerte representación de los adultos más jóvenes en el mercado informal *vis-à-vis* los demás factores. La explicación de tales resultados desborda a los modelos presentados, podrían ser involuntarias, o voluntarias (para el caso de los jóvenes adultos, por ejemplo), pero constituyen un factor a tener en cuenta en el diseño de política.

Referencias

Amemiya, T. (1985). *Advanced Econometrics*.

Blanchflower, D. G. and Oswald, A. J. (1989). The wage curve. Technical report, National Bureau of Economic Research.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Card, D. (1995). The wage curve: a review.

Nelson, C. and Startz, R. (1988). The distribution of the instrumental variables estimator and its t-ratiowhen the instrument is a poor one.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415.

Sargan, J. D. (1988). Testing for misspecification after estimating using instrumental variables. *Contributions to Econometrics: John Denis Sargan*, 1:213–235.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.