Capítulo de Tesis de Doctorado.

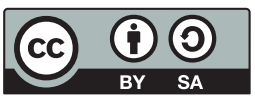
Anexo 3. Sesgo de Selección. Efectos de Origen de Clase en Argentina (1955-2001).

Quartulli, Diego.

Cita:

Quartulli, Diego (2016). *Anexo 3. Sesgo de Selección. Efectos de Origen de Clase en Argentina (1955-2001)*. Capítulo de Tesis de Doctorado.

Dirección estable: **https://www.aacademica.org/diego.quartulli/53**



Anexo 3
Sesgo de selección

El diseño de una muestra no puede, posiblemente,
ser planeado para todas las inferencias de poblaciones
que podrían ser hechas por los investigadores y por los lectores
(Kish, 2004, p. 30)

En las situaciones que existe buenas razones para suponer la presencia de un sesgo sistemático que, *a posteriori*, produzca inferencias sesgadas también existen buenas razones para implementar algún tipo de ajuste por imperfecto que este sea (Little & Rubin, 1987).

Siguiendo esta premisa, en lo que sigue se intentará primero localizar y luego corregir (con limitaciones), el *sesgo de selección* derivado del *diseño muestral* con la que se trabajará.[1]

Como se detalló en la sección 'Estructura de datos' del capítulo 4 (§4.2), el mayor sesgo de un *diseño transversal con preguntas retrospectivas* es la extinción real de las unidades de análisis en la población a medida que se aleja retrospectivamente hacia el período de referencia de las preguntas (Winship & Mare, 1992)(Noymer, 2001).

Este tipo de problemas, que se podría denominar de *casos faltantes*, se suelen atenuar mediante una calibración basada en modelos teóricos y difiere cualitativamente de aquellos que se producen por *datos faltantes* (*missing data*) que son usados, por ejemplo, en las estimaciones de ingreso.[2]

En términos generales podría suponerse que los problemas analizados anteriormente, esto es:

[1] Cabe aclarar que la muestra ya tiene sus ponderadores y expansores aunque calibrados para la población efectiva de 2010 en función del diseño muestral. El ponderador sobre el sesgo de selección se aplica en forma posterior a los anteriores. Para más detalles puede consultarse el anexo metodológico y la siguiente bibliografía externa propia de la EDSA (Quartulli, Tinoboras, Vera, & De Grande, 2011)(Quartulli, 2012).

[2] Para los primeros tipos de problemas puede consultarse el clásico libro de Leslie Kish (Kish, 1965). Para versiones más actuales y flexibles puede consultarse la obra de Särndal (Deville & Särndal, 1992)(Särndal, 2007). Para el segundo tipo de problemas puede consultarse la clásica introducción de Little y Rubin (Little & Rubin, 1987).

Para una aplicación al caso argentino, en donde se analizan los potenciales sesgos de la no respuesta de ingresos, su variación en el tiempo, las diferentes técnicas de imputación y se propone una solución aproximada puede consultarse (Donza, 2011)(Donza, 2013).

a) el de los casos faltantes debido a la mortalidad y
b) su distribución no aleatoria en función de las variables de interés a analizar

podrían representarse de forma muy abstracta, al menos para datos de tablas de contingencia, como lo sugiere la expresión A3.1:

A3.1

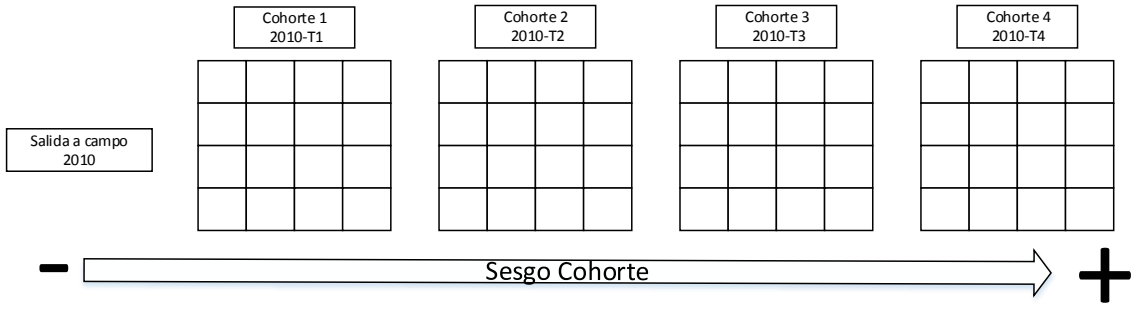
$$M_i^j = \beta coh_j \oplus \beta cel_i$$

Donde M_i^j es la mortalidad de la cohorte j y la celda i y M_i^j se considera igual a la agregación del efecto del sesgo específico de la cohorte ($\boldsymbol{\beta coh_j}$) y del efecto específico de la celda ($\boldsymbol{\beta cel_i}$). La expresión A3.1 no especifica el modo en que se deberían agregar ambos efectos ni que designan los símbolos utilizados para nombrar a los efectos cohorte y celda. Aun así, es útil para dar una aproximación intuitiva a lo que se quiere expresar.

A3.1 Sesgo Cohorte y Sesgo Celda

Suponiendo que las relaciones intergeneracionales a analizar de cada cohorte se puedan representar como una tabla de contingencia y que se distinguen 4 categorías en origen y 4 en destino, parece válido el supuesto que asume que aquella mortalidad es creciente en función del tiempo y que $T1 < T2 < T3 < T4$. Esto se intenta representar en la Figura A3.2.

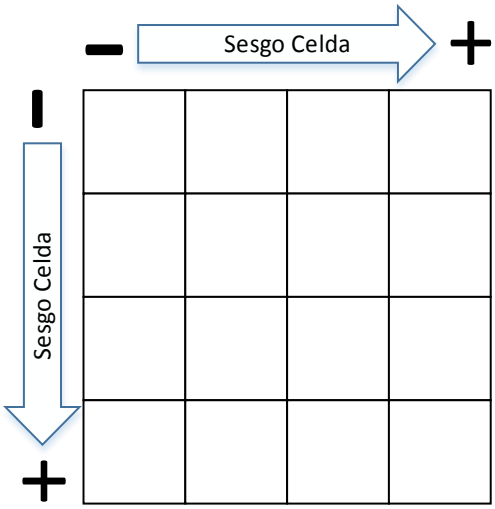
Figura A3.2. Esquematización del Sesgo Cohorte.



La interpretación de este sesgo se puede dilucidar como una gran media de la mortalidad que sirve para estimar la magnitud de la misma. Obsérvese que si esta fuera aleatoria en función de otras características sociales, este podría ser un sesgo ignorable ya que no afectaría al tipo de análisis *inter-cohortes* que se desarrollan en los capítulos empíricos.

El *sesgo celda* puede representarse, como lo siguiere la Figura A3.3, como un sesgo *intra-cohorte*. Este valor, debido a los objetivos de la investigación, es importante ya que implica que algunas de las diferencias encontradas en los análisis de flujos relativos puedan deberse a este tipo de sesgo. Para hacer perceptible el argumento se asumirá que los orígenes y los destinos son ordinales y que crecen de izquierda a derecha y de arriba hacia abajo.

Figura A3.3. Esquematización del Sesgo Celda.



Si se piensa que los datos a analizar son categóricos y los mismos se computan como frecuencias de las celdas de una tabla de contingencia, la expresión A3.1 junto con las Figuras A3.2 y A3.3 puede pensarse como una mortalidad específica para cada celda. En este sentido, la intersección de un origen y un destino de cada tabla de contingencia (cada cohorte) debería tener un valor específico. Esta diferencia, sí constituye un *sesgo de selección*, ya que concentra mucho de la mortalidad de toda la muestra en las celdas del rincón inferior derecho de todas las cohortes, pero muy en especial de la cohorte más vieja.

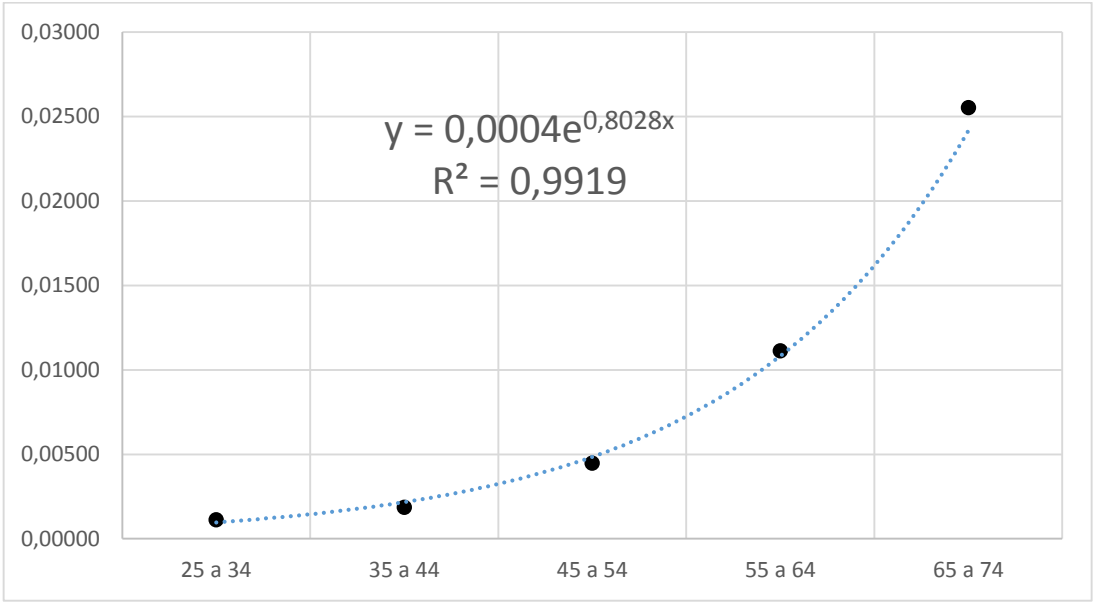
A3.2 Construcción del Ponderador

Aclarado en términos intuitivos la expresión A3.1, se pasa a describir las acciones que se realizaron tendientes a estimar y atenuar los sesgos anteriores.

El *sesgo cohorte*, esto es, cuantos individuos menos se encontraron en cada cohorte, se podría estimar con la calibración de un modelo de la mortalidad de cada individuo según la edad. Se admite el supuesto que al menos para las edades que se intenta imputar (+25 años), la mortalidad no sólo es una función monótona creciente, sino que la misma se puede modelar aceptablemente como una función exponencial de la edad (Keyfitz & Caswell, 2005).[3]

La Figura A3.4 aporta información *externa* que corrobora razonablemente el supuesto anterior y especifica que dentro del conjunto de funciones monotónicas crecientes, la tasa de mortalidad adulta se puede razonablemente representar como a una función exponencial.[4]

Figura A3.5. Tasa específica de mortalidad (y) por grupo de edad (x) para Argentina 2010. Bondad de Ajuste del modelo según coeficiente de correlación de Pearson (R^2)



Elaboración propia en base a Estadísticas Vitales 2010. Especialmente en base a cuadros 22, 23a y 23b (Ministerio de Salud de la República Argentina, 2011, pp. 25-33).

[3] Este supuesto no se cumpliría si se intentara imputar la mortalidad desde el nacimiento biológico de los individuos. Este tendría una forma de una "$\mathcal{J}$", debido a la mortalidad infantil y su tratamiento matemático sería algo más complicado.

[4] Si se supone, como hipótesis de trabajo, que la mortalidad específica no hay variado durante el tiempo calculado, es interesante notar que sigue quedando en pie mucho de la exposición original de Leonard Euler en 1760 sobre las tablas de mortalidad (Euler, 1767 [1760]).

A pesar de existir datos sobre edades más desagregadas, para el momento de averiguar la bondad de ajuste del modelo y la parametrización de su correspondiente función, se prefirió trabajar con datos algo más agregados ya que de este modo se reducen los errores de medición y las estimaciones resultantes pueden considerarse más robustas.

Dado que este modelo, a pesar de contener pocos parámetros y cierta elegancia, ajusta de forma precisa, luego se lo usó como modelo teórico para estimar la cantidad de individuos que deberían haber entrado en la investigación y no lo hicieron por su respectiva mortalidad.[5]

El *sesgo celda*, esto es, la desigualdad de la mortalidad según distintos factores de heterogeneidad presente en la población, se estimó con un proxy construido sobre datos de la diferencial tasa de mortalidad según nivel educativo y el sexo. Esta decisión, se debe tanto a la disponibilidad *externa* de fuentes secundarias que incluyan esas variables como a la disponibilidad *interna* de ellas en el cuestionario de la muestra a corregir. Diversas investigaciones internacionales hacen plausible la confiabilidad de este proxy.[6] [7]

En este sentido, gracias a la disponibilidad de datos sobre edad y sexo de *forma conjunta* en las estadísticas vitales, es posible calcular un modelo exponencial para cada uno de esas poblaciones.

---

[5] Para fijar las ideas acerca de la magnitud de este proceso para el período en cuestión (1955-2010) de los aproximadamente 4000 individuos que conformaban la muestra, se pueden interpretar como los sobrevivientes de una población de 5000 (aprox.) individuos mayores a 26 años. Se toma hasta 2010 porque al ser el año en que se realizó la muestra, hasta ese año tendrían que haber llegado los sobrevivientes del universo a indagar.

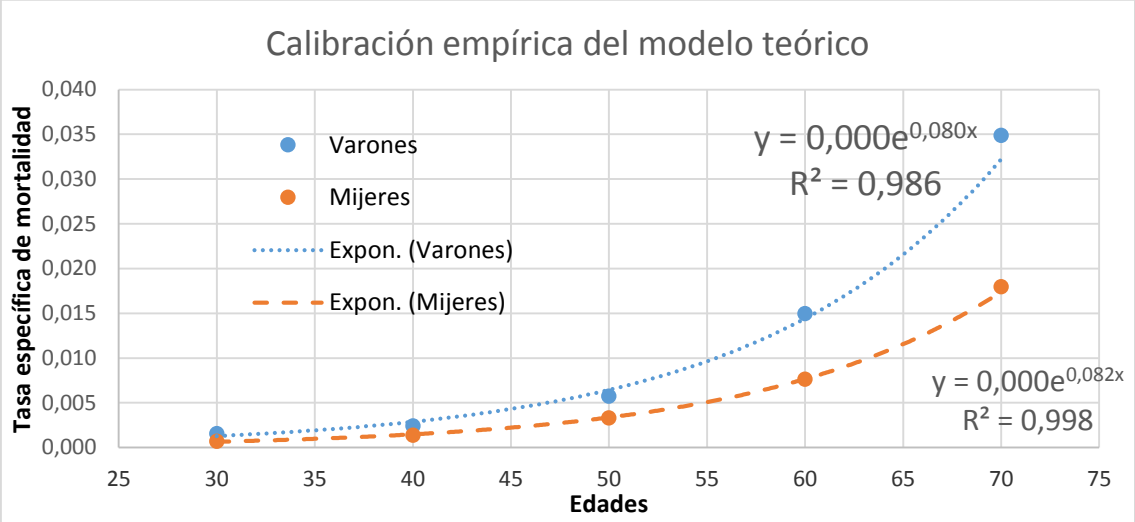
[6] En lo referido a la disponibilidad externa a nivel internacional puede consultarse el trabajo de Shkolnikov *et. al.* Este demuestra la existencia de diferenciales de mortalidad según nivel educativo como su cambiante tendencia según países (Shkolnikov, Andreev, Jasilionis, Leinsalu, & Antonova, 2006).

En cuanto a la relación entre la mortalidad y una serie de indicadores sobre posición social (como clase, status, educación, ingreso, etc.) puede consultarse el excelente trabajo de Torssander y Erikson (Torssander & Erikson, 2008).

[7] Para el caso Argentino, existe un trabajo de Peláez y Acosta para la provincia de Córdoba, en donde también se constatan diferenciales en las tasas de mortalidad según nivel educativo. El trabajo de Peláez y Acosta es interesante por intentar estimar, de manera ingeniosa, la influencia del nivel educativo en la mortalidad de Argentina, en ausencia de datos oficiales de las estadísticas vitales. Ellos, estiman los datos siguiendo una cohorte teórica a través de diferentes censos nacionales de población, en presencia de supuestos como el predominio de una sociedad cerrada.

Si bien a nivel desagregado los resultados no son muy alentadores, a un nivel más agregado sus trabajos permiten una estimación razonable de la influencia del nivel educativo en la mortalidad adulta (Peláez & Acosta, 2011).

Figura A3.6. Tasa específica de mortalidad (y) por grupo de edad (x) para Argentina 2010. Bondad de Ajuste del modelo según coeficiente de correlación de Pearson (R^2). Los datos se encuentran redondeados a 3 decimales.



Luego, cada una de las respectivas curvas exponenciales fueron corregidas por un coeficiente que tiene en cuenta el impacto diferencial del nivel educativo. Nuevamente, para reducir errores de medición y construir estimaciones más robustas, se estimó un coeficiente que corrija sólo para dos grandes niveles educativos como 'Hasta secundaria incompleta' y 'Secundaria completa y más'.

Esto se hizo cotejando los datos que arrojaron los diferentes censos de población desde 1960 hasta 2001, tomando como valor a corregir el valor del 2001 en función de los valores de los censos anteriores. Esta estrategia proyectó estimaciones en el sentido de lo esperado, indicando que cuanto más vieja era la cohorte a analizar, más había que sobre-representar a su población con "Hasta secundaria incompleta" y sub-representar a su población con "Secundaria completa y más".

Tomando un promedio de todos los censos, ya que el factor temporal clave lo capta la parte exponencial del ponderador, se llega a las expresiones A3.2 y A3.3 que logran condensar todo lo anterior y relacionar de manera precisa lo que en forma intuitiva se había representado en la expresión A3.1:

A2.2

$$M_{muj} = ae^{\beta x} * M_{ed}$$

A3.3

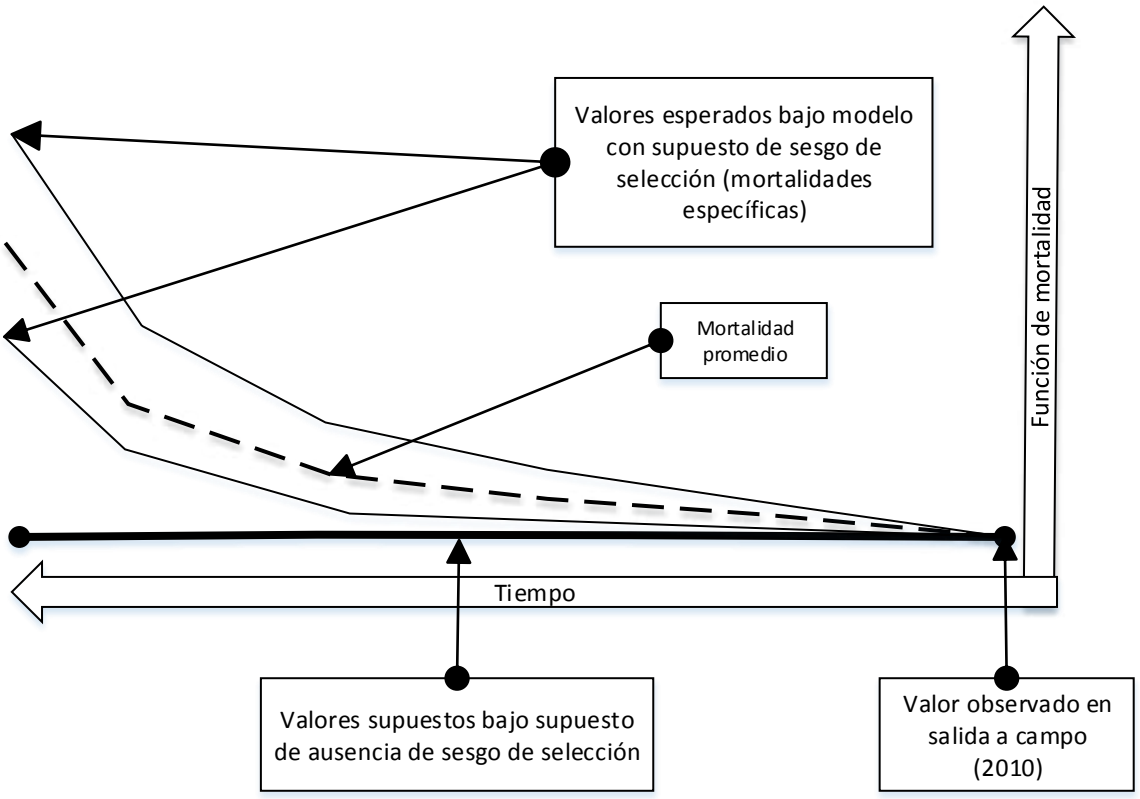
$$M_{hom} = ae^{\beta x} * M_{ed}$$

En donde cada expresión se aplica a una población diferente (A3.2 para mujeres y A3.3 para hombres), $ae^{\beta x}$ es la parte exponencial en donde tanto el intercepto a como el parámetro β se calibran empíricamente (Figura A3.6), e se supone la

constante de Euler, x son los años y M_{ed} es la mortalidad según nivel educativa que, se supone a efectos del presente cálculo indiferente según sexo.

Lo importante de este ponderador es que, por más imperfecto que pueda ser, logra corregir parte del *sesgo de selección* producido por la mortalidad diferencial. Que tan cerca queda el *estimador muestral* del verdadero *parámetro poblacional* es una incógnita. Lo que sí se puede afirmar, es que con las operaciones efectuadas ambos valores se encuentran más cerca uno del otro, en comparación, a la inacción derivada de suponer que no existe un sesgo de selección. Este razonamiento se visualiza en la Figura A3.7.

Figura A3.7. Efecto diferencial del sesgo de selección para dos poblaciones diferentes, manteniendo su característica exponencial.



Obviamente, los problemas anteriores, relacionados con el tipo de *diseño de la investigación* y su correspondiente *estructura de datos*, no deben confundirse con problemas propios de los *análisis de los datos*.