UNIVERSIDAD SANTO TOMÁS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Estimador de impacto por variables instrumetales calibradas

Deisy Yomaira Devia R.

Universidad Santo Tomas
Division de ciencias economicas
Facultad de Estadística
Bogotá D.C., Colombia
2016

Estimador de impacto por variables instrumetales calibradas

Deisy Yomaira Devia R.

Trabajo de grado presentado como requisito parcial para obtar al título:

Profesional en Estadística

Director:
Ph.D. Hugo Andrés Gutiérrez Rojas

Universidad Santo Tomas
Division de ciencias economicas
Facultad de Estadística
Bogotá D.C., Colombia
2016

DEDICATORIA

A Dios por iluminar mi camino y enseñarme que la vida no es casual sino causal, todo tiene su tiempo y razón de ser, es así como haz guiado mis pasos para por fin alcanzar esta meta. A mis seres queridos por su apoyo y constante motivación. También a los docentes por su paciencia y guía en este trayecto.

Agradecimientos

Agradezco a Dios, a mi familia por su apoyo incondicional, a mis maestros de la Universidad Santo Tomas y en especial a mi director H. Andrés Gutiérrez R. a quien admiro y respeto, le agradezco por su acompañamiento y paciencia en el asesoramiento del trabajo de grado, pues a pesar de sus muchas obligaciones ha hecho espacio para colaborarme.

Resumen

En este trabajo se propone un estimador calibrado para la evaluación de impacto por variables instrumentales, ya que, como se ha comprobado en distintas investigaciones esta técnica es robusta y eficiente. Vía simulación se logró determinar que el estimador propuesto obtiene estimaciones con sesgos tan pequeños o menores que los obtenidos mediante el estimador clásico. Adicional a esto bajo escenarios específicos en donde no se cumplen algunas condiciones y se tienen pequeños tamaños muestrales, se obtuvo que el estimador calibrado propuesto es más eficiente comparado con el estimador clásico mostrando de este modo que es una opción factible para aplicar en la evaluación de impacto.

Palabras clave: calibración, causalidad, evaluación de impacto, variables instrumentales.

Abstract

This paper is proposed a calibrated estimator for impact evaluation by instrumental variables, because as has been proven in various studies this technique is robust and efficient . Through simulations is determined that the calibrated estimator gets estimations equally biased or less than those obtained by the classical estimator. additional to this under specific scenarios where certain conditions are not met and is have small sample sizes, it was found that the proposed calibration estimator is more efficient compared to the classical estimator thus showing that it is a feasible option for apply in assessing impact.

Keywords: calibration, causality, impact assesment, instrumental variables.

Índice general

Capítulo 1

Motivación

Distintas organizaciones gubernamentales reconocen la necesidad de que los procesos de toma de decisiones, creación de políticas públicas y evaluación de estas mismas sean más efectivas. Ya que se concibe el gobierno como un instrumento para el planteamiento y desarrollo de programas de política pública, en su gestión por dar solución a los problemas colectivos que se presentan en el estado.

Se hace por lo tanto necesario basar en sólidas investigaciones y evidencias los resultados que se generan por estas intervenciones sobre los participantes, es decir que se busca saber cuál es el efecto causal del programa, logrando de esta manera mejorar la efectividad y funcionamiento. Por lo tanto bajo esta premisa se han desarrollado distintas metodologías estadísticas como son el diseño experimental aleatorio, el matching, el propensity score matching, el método de diferencias en diferencias, el diseño de regresión discontinua y el método de variables instrumentales. Las cuales se implementan dependiendo del tipo de evaluación y como se ha desarrollado esta.

En la actualidad con los distintos objetivos de desarrollo en el estado a nivel gubernamental y social se requiere que estas herramientas de evaluación sean cada día más eficaces y permitan obtener estimaciones lo más acertadas posibles. Es por esto que en el que hacer del estadístico esta proponer nuevas metodologías y seguir mejorando las ya existentes. Por lo cual nos centramos en este trabajo de investigación en desarrollar estimaciones

más efectivas por medio de la metodología de variables instrumentales aplicando pesos de calibración por el enfoque funcional, estos métodos no sólo son muy utilizados en esta área sino que además son útiles en otros campos como Econometría, Epidemiologia y afines.

El Objetivo General de esta investigación es desarrollar una metodología que incorpore dentro del método de variables instrumentales pesos de calibración por el enfoque funcional generando estimadores de impacto más precisos. Así mismo se tienen como objetivos específicos 1) Revisión de la metodología para la evaluación de políticas públicas por variables instrumentales. 2) Proponer una expresión para el estimador del impacto de un programa con pesos calibrados. 3) Proponer expresiones para la esperanza y la varianza de ese estimador. 4) Estudiar empíricamente el estimador por simulaciones. 5) Aplicación en una situación real.

Capítulo 2

Revisión

2.1. Evaluación de políticas públicas

(Departamento Nacional de Planeación [DNP]. SINERGIA, 2012) nos introducen en la evaluación de política pública, definiéndola como la investigación sistemática dedicada a medir analizar y determinar la significación o valor de una intervención pública sea política, norma, programa o plan. La cual es aplicada en los eslabones de la cadena de resultados (Insumos, procesos, productos, resultados e impactos), con el fin de ser útil a los tomadores de decisiones y gestores públicos, logrando de este modo mejorar el proceso de diseño, implementación y efectos de estas intervenciones.

Existen diferentes tipos de evaluaciones como son

Evaluación de procesos Permite identificar las relaciones y actividades que se requieren dentro de los procedimientos y tareas de la implementación del programa, para producir el bien o servicio que se desea de la intervención.

Evaluación de productos Analiza en qué medida se ha logrado maximizar los productos con el menor costo o por que no se han obtenido los productos planteados.

Evaluación de resultados Estudia cambios que se dieron debido a los

productos de la intervención en los beneficiarios sea directa o indirectamente.

Evaluación costo-beneficio Realiza comparaciones de costos y beneficios de una intervención asignando valores monetarios, permitiendo estandarizar los resultados y la distribución de recursos de forma eficiente.

Evaluación institucional Busca medir la eficiencia de los entes o instituciones que se encargan de generar e implementar el programa o política.

Evaluación de impacto Este tipo de evaluación se realiza después de implementar el programa permitiendo medir y evidenciar si es atribuible a la intervención los efectos generados sobre la población beneficiaria.

En este trabajo nos enfocamos en este último tipo de evaluación, en la cual se genera escenarios cuasi-experimentales y buscamos conocer el efecto causal del programa.

2.2. Causalidad

En la literatura actual (Morgan y Winship(2007), D. Angrist y Steffen (2008), García (2010) y Cacheiro (2011)), se considera para la medición del efecto causal una amplia teoría de inferencia causal estadística, la cual permite tener determinado control sobre los distintos factores que puedan afectar la variable de interés. De tal manera que, las condiciones en toda la población sean idénticas y soló difieran en la intervención, de este modo el impacto será atribuible a el programa.

2.2.1. Modelo causal de Neyman-Rubin

Consideramos ahora el modelo propuesto por (Neyman(1990) y Rubin(1974) citado en Morgan y Winship(2007), D. Angrist y Steffen (2008), García (2010)) denominado modelo causal de Neyman-Rubin, donde la variable

de interés, Y_i es escrita en términos de posibles respuestas para cada individuo de la población.
De modo que, se tiene una variable T_i que indica la exposición

$$T_i = \begin{cases} 1 & \textit{si el individuo es expuesto al tratamiento} \\ \\ 0 & \textit{si el individuo es expuesto al control} \end{cases}$$

También se tiene la variable de interés Y_i, dada la exposición en el individuo en términos de resultados observables o no observables.

$$Y_i(0) = \begin{cases} y_i^{no-obs} & si \quad T_i = 1 \\ \\ y_i^{obs} & si \quad T_i = 0 \end{cases}$$

$$Y_i(1) = \begin{cases} y_i^{no-obs} & si \quad T_i = 0 \\ \\ y_i^{obs} & si \quad T_i = 1 \end{cases}$$

Dadas las anteriores expresiones, se tiene que el impacto del programa es dado por la diferencia entre lo ocurrido con el individuo en presencia del programa y lo ocurrido con el mismo individuo en ausencia del programa.

$$\alpha = y_i^{obs} - y_i^{no-obs}$$

Este escenario es definido como contrafactual en la práctica no es observable, por lo que se debe estimar por medio del grupo control.

Según lo planteado anteriormente se tienen los siguientes efectos (α) :

- Efecto medio del tratamiento

$$\alpha_{ATE} = E\left[Y_i(1) - Y_i(0)\right]$$

- Efecto del tratamiento sobre los tratados

$$E\left[Y_i \mid T = 1\right] - E\left[Y_i \mid T = 0\right] = \underbrace{E\left[Y_i(1) - Y_i(0) \mid T = 1\right]}_{\alpha_{ATT}}$$

$$+ \underbrace{E\left[Y_i(0) \mid T = 1\right] - E\left[Y_i(0) \mid T = 0\right]}_{Sesgo\ de\ selección}$$

- Efecto del tratamiento sobre los no tratados

$$E\left[Y_i \mid T=1\right] - E\left[Y_i \mid T=0\right] = \underbrace{E\left[Y_i(1) - Y_i(0) \mid T=0\right]}_{\alpha_{ATC}}$$

$$+ \underbrace{E\left[Y_i(1) \mid T=1\right] - E\left[Y_i(1) \mid T=0\right]}_{Sesgo\ de\ selección}$$

Finalmente para estos dos últimos efectos se busca controlar el sesgo de selección, para lo cual se supone y valida que:

- Hay independencia condicional
- No hay predictibilidad perfecta para los individuos que se benefician o no del programa

Este estimador puede ser obtenido por mínimos cuadrados ordinarios o máxima verosimilitud de la expresión:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \tag{2.1}$$

En la cual el estimador $\hat{\beta}_1$ será la estimación del impacto del programa. Pero en la práctica el término de sesgo de selección no es cero, ya que la selección de los individuos que son beneficiarios del programa se hace bajo ciertos criterios de focalización entre otros. Por ende se recurre a métodos que permiten quitar el sesgo de selección, dentro de los cuales se encuentra el de variables instrumentales.

2.3. Estimación de coeficientes de regresión

Como se plantea en Gutiérrez (2009), se supone la existencia de las variables aleatorias $Y_1, Y_2, \cdots, Y_N$ y del vector de variables aleatorias $X_1, X_2, \cdots, X_N$, que están relacionados por el modelo de probabilidad ξ dado como

$$Y_k = X_k^{\iota}\beta + \varepsilon_k \tag{2.2}$$

Donde β es el vector de coeficientes de regresión para el modelo; ε_k representan los términos de error del modelo, que son variables aleatorias independientes e idénticamente distribuidas, con media cero y varianza $c_k\sigma^2$.

Dado esto se tiene que

$$E_\varepsilon(Y_k) = X_k^{'}\beta \tag{2.3}$$

$$V_\varepsilon(Y_k) = c_k\sigma^2 \tag{2.4}$$

Sin embargo no se puede conocer β, ya que dichas variables aleatorias no son observables en su totalidad. Por lo tanto se tienen los valores $y_1, y_2, \cdots, y_N$ que son realizaciones del vector de variables aleatorias $Y_1, Y_2, \cdots, Y_N$ e igualmente para el vector $X_1, X_2, \cdots, X_N$, es decir que se cuenta con una población finita de tamaño N. Donde las respectivas realizaciones están relacionadas por el modelo poblacional:

$$y_k = x_k^{'}B + E_k \tag{2.5}$$

Conforme al modelo anterior se obtiene el estimador B por el método de mínimos cuadrados, que resulta de minimizar la siguiente función

$$D = \sum_u \left(\frac{y_k - x_k^{'}B}{c_k\sigma^2} \right)^2 \tag{2.6}$$

Al minimizar D se obtiene la siguiente expresión para B

$$B = \left(\sum_u \frac{x_k x_k^{'}}{c_k\sigma^2} \right)^{-1} \sum_u \frac{x_k y_k}{c_k\sigma^2} \tag{2.7}$$

$$= \left(\sum_u \frac{x_k x_k^{'}}{c_k} \right)^{-1} \sum_u \frac{x_k y_k}{c_k} \tag{2.8}$$

$$= T^{-1}t \tag{2.9}$$

En la práctica no se conoce para cada k de la población finita los valores de la variable de interés, ni de la información auxiliar, por lo que B debe ser estimado.

Por lo tanto la estimación de B por el método de mínimos cuadrados será

$$\hat{B} = \hat{T}^{-1}\hat{t} \tag{2.10}$$

$$\hat{T}^{-1} = \sum_s \frac{x_k x_k^{'}}{\pi_k c_k} \tag{2.11}$$

$$\hat{t} = \sum_s \frac{x_k y_k}{\pi_k c_k} \tag{2.12}$$

Este estimador no es insesgado, pero igualmente (Sarndal, Swensson y Wretman (1992)citado en Gutiérrez (2009)) dan una aproximación de la varianza

$$AV(\hat{B}) = \left(\sum_u \frac{x_k x_k^{'}}{\sigma_k^2} \right)^{-1} V \left(\sum_u \frac{x_k x_k^{'}}{\sigma_k^2} \right)^{-1} \tag{2.13}$$

Donde V es una matriz simétrica y sus entradas están dadas por

$$v_{ij} = \sum \sum_u \Delta_{kl} \left(\frac{x_{ik} E_k}{\pi_k} \right) \left(\frac{x_{jl} E_l}{\pi_l} \right) \tag{2.14}$$

Con $E_k = y_k – x_k^{'} B$ La estimación estará dada por

$$\hat{Var}(\hat{B}) = \left(\sum_s \frac{x_k x_k^{'}}{\sigma_k^2 \pi_k} \right)^{-1} \hat{V} \left(\sum_s \frac{x_k x_k^{'}}{\sigma_k^2 \pi_k} \right)^{-1} \tag{2.15}$$

Las entradas de la matriz simétrica estimada $\hat{V}$ son

$$\hat{v}_{ij} = \sum \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{x_{ik} e_k}{\pi_k} \right) \left(\frac{x_{jl} e_l}{\pi_l} \right) \tag{2.16}$$

Con $e_k = y_k – x_k^{'} \hat{B}$

2.4. Pseudo-verosimilitud

Otro método para la estimación de coeficientes de regresión es el de máxima verosimilitud, el cual, se emplea bajo algunos supuestos sobre la distribución del término de error. Siguiendo a Gutiérrez (2014), consideramos el modelo de super-población ξ

$$Y_k = \beta_0 + \beta_1 X_1 + \varepsilon_k \tag{2.17}$$

donde

$$\varepsilon_k \sim N(0, \sigma^2) \qquad k = 1, \cdots, N \tag{2.18}$$

La anterior expresión es lo mismo que decir

$$Y_k \sim N(X_k^{'}\beta, \sigma^2) \qquad K = 1, \cdots, N \tag{2.19}$$

Como se mencionó anteriormente no es posible conocer β. Por ende dadas las respectivas realizaciones y_k y x_k de las variables aleatorias Y_k y X_k, suponemos el siguiente modelo poblacional

$$y_k = B_0 + B_1 x_1 + E_k \tag{2.20}$$

Para el cual, consideramos la función de densidad poblacional $f(y; B)$, de manera que, el estimador de máxima verosimilitud para los parámetros de interés B se obtiene al maximizar con respecto a este la función máximo verosímil

$$L(y_1, y_2, \cdots, y_N; B_0, B_1, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{k\in U} y_k{-}(B_0 - B_1 x_k)^2\right) \tag{2.21}$$

$$l(y_1, y_2, \cdots, y_N; B_0, B_1, \sigma^2) = -\frac{N}{2}log(2\pi){-}\frac{N}{2}log(\sigma^2){-}\frac{1}{2\sigma^2}\sum_{k\in U} y_k{-}(B_0{-}B_1 x_k)^2 \tag{2.22}$$

Al derivar la anterior función con respecto al parámetro de interés se obtiene la función de score o ecuaciones de verosimilitud poblacionales $U(y, B)$

$$U(y, B) = \frac{\partial l(y_1, y_2, \cdots, y_N; B_0, B_1, \sigma^2)}{\partial B} = 0 \tag{2.23}$$

Donde cada score estará dado por

$$u_k = \frac{\partial log f(y, B)}{\partial B} \tag{2.24}$$

De las anteriores ecuaciones se obtendrá el estimador B_U para la población finita, pero en la práctica no se tiene acceso a toda $k \in U$. Por lo tanto, algunos autores, (Fuller (2009), Chambers & Skinner (2003, p. 179) y Pessoa & Silva(1998, capítulo 5), citado en Gutiérrez, 2014), han planteado

el enfoque de pseudo-verosimilitud, el cual, permite bajo la información recopilada en la muestra, tomar a $\sum_{k\in U} u_k(\beta)$ como un parámetro que se desea estimar. Para esto se aplica el estimador Horvitz-Thompson, ya que este es un estimador lineal ponderado, de modo que

$$\sum_{k\in s} d_k u_k(B) \tag{2.25}$$

Donde $d_k = \frac{1}{\pi_k}$.

Dado esto podemos obtener el estimador $\hat{B}_s$ al resolver las ecuaciones de pseudo-verosimilitud dadas por

$$\sum_{k\in s} d_k u_k = 0 \tag{2.26}$$

Para este estimador su varianza estará dada por

$$V(\hat{B}_s) = [J(B_U)]^{-1} V\left[\sum_{k\in s} d_k u_k\right] [J(B_U)]^{-1} \tag{2.27}$$

Donde $V\left[\sum_{k\in s} d_k u_k\right]$ representa la varianza del diseño muestral del estimador de scores y $J(B_U)$ es la información de Fisher dada por

$$J(B_U) = \frac{\partial \sum_{k\in U} u_k(B_U)}{\partial(B_U)}|_{B=B_U} \tag{2.28}$$

La estimación de esta varianza será

$$\hat{V}(\hat{B}_s) = \left[\hat{J}(\hat{B}_s)\right]^{-1} \hat{V}\left[\sum_{k\in s} d_k u_k(\hat{B}_s)\right] \left[\hat{J}(\hat{B}_s)\right]^{-1} \tag{2.29}$$

Donde $\hat{V}\left[\sum_{k\in s} d_k u_k(\hat{B}_s)\right]$ es la varianza estimada por el diseño muestral para el total poblacional de scores y $\hat{J}(\hat{B}_s)$ estará dado por

$$\hat{J}(\hat{B}_s) = \frac{\partial \sum_{k\in s} d_k u_k(\hat{B})}{\partial B}|_{B=\hat{B}_s} \tag{2.30}$$

2.5. Calibración

Como lo explican algunos estadísticos (Sarndal & Deville (1998); Estevao, Sarndal & Sautory (2000), citado en Gutiérrez, 2009), esta metodología surge de la incorporación de información auxiliar, mejorando de este modo las estimaciones de la encuesta. Por lo tanto al disponerse de información auxiliar correlacionada con la variable de interés y conocerse los totales de estas a nivel poblacional. Se busca generar un sistema de pesos w_k, para cada k que pertenece a la muestra s y dependen de la información auxiliar. Además estos deben cumplir con:

$$\sum_{k \in s} w_k x_k = t_x \tag{2.31}$$

- Que hace referencia a la propiedad de consistencia, reproduciendo el total conocido para cada variable auxiliar.
- Cercanía a los pesos básicos es decir estos son tan cercanos como sea posible a $d_k = \frac{1}{\pi_k}$. Induciendo de este modo estimaciones aproximadamente o asintóticamente insesgadas.

Estos pesos pueden construirse por dos enfoques:

2.5.1. Minimización de distancia

Este consiste en minimizar entre w_k y d_k una pseudo distancia, que cumple con las condiciones mencionadas, además de los siguientes supuestos que se mencionan en Gutiérrez (2009)

- Es estrictamente no negativa
- Es estrictamente convexa
- Distancia entre pesos iguales es cero
- La función tiene un punto crítico cuando los pesos son iguales
- El minimizador corresponde al punto $G^{'}(1) = 0$

Se realiza una optimización por multiplicadores de lagrange y se define una función $F(\cdot)$, la cual permite llegar a los pesos

$$w_k = a_k F(q_k \lambda^{`} x_k) \tag{2.32}$$

Obteniendo λ por la solución del sistema de ecuaciones

$$\sum_{k \in s} a_k F(q_k \lambda^{`} x_k) x_k^{`} = t_x^{`} \tag{2.33}$$

2.5.2. Enfoque funcional

Sarndal & Estevao (2000) plantean este enfoque, bajo el cual, los pesos de calibración se construyen a partir del vector auxiliar x_k, donde el vector de totales $t_x = \sum_u x_k$ es conocido. Además se define una constante positiva q_k y el vector de instrumentos z_k para todo k en la muestra. Estos parámetros satisfacen las siguientes condiciones

- $dim(z_k) = J = dim(x_k)$
- $\sum_s q_k z_k x_k^{`}$ es una matriz no singular

Dado lo anterior se generan los pesos calibrados

$$w_k = (d_k + q_k \lambda^{`} z_k) \tag{2.34}$$

Donde λ es un vector que depende de s pero no de k y está dado por

$$\lambda = \left(t_x - \hat{t}_{x\pi}\right) \left(\sum_{k \in s} q_k z_k x_k^{`} \right)^{-1} \tag{2.35}$$

Para el cual $\hat{t}_{x,\pi}$ es el vector de totales estimados por Horvitz-Thompson.

2.5.3. Estimación con pesos calibrados enfoque funcional

Con este sistema de pesos el estimador de calibración estará dado por

$$\hat{t}_{calf} = \sum_{k \in s} w_k y_k \tag{2.36}$$

$$= \sum_{k \in s} d_k y_k + \sum_{k \in s} q_k \lambda_k z_k y_k \tag{2.37}$$

$$= \hat{t}_y + \left(t_x - \hat{t}_{x\pi}\right)^{'} R_k \sum_{k \in s} y_k \tag{2.38}$$

Donde

$$R_k = \left(\sum_{k \in s} q_k z_k x_k^{'}\right)^{-1} \left(\sum_{k \in s} q_k z_k\right) \tag{2.39}$$

Este estimador no es insesgado pero su sesgo es pequeño como lo muestran (Sarndal et,al. ,2000) e igualmente dan una aproximación de la varianza y su varianza estimada

$$AV\left(\hat{t}_{calf}\right) = \sum \sum_{u} D_{kl} H_k H_l \tag{2.40}$$

Donde $D_{kl} = \frac{d_k d_l}{d_{kl}-1}$, $\quad d_{kl} = \frac{1}{\pi_{kl}}$ para $k \neq l$

Para $k = l \qquad D_{kl} = d_k - 1$

$$H_k = E_k + L_k = y_k - x_k^{'} B + x_k (B - Q) \tag{2.41}$$

$$= y_k - x_k^{'} Q \tag{2.42}$$

Con

$$E_k = y_k - x_k^{'} B \tag{2.43}$$

$$B = \left(\sum_{u} \frac{x_k x_k^{'}}{c_k}\right)^{-1} \sum_{u} \frac{x_k y_k}{c_k} \tag{2.44}$$

$$Q = \left(\sum_{u} \pi_k q_k z_k x_k^{'}\right)^{-1} \sum_{u} \pi_k q_k z_k y_k \tag{2.45}$$

En cuanto a la estimación de la varianza esta será

$$\hat{V}(\hat{t}_{calf}) = \sum \sum_{s} D_{kl} h_k h_l \tag{2.46}$$

Con

$$h_k = y_k - x_k^{'} \hat{Q} \tag{2.47}$$

$$h_l = y_l – x_l^{'} \hat{Q} \tag{2.48}$$

$$\hat{Q} = \left(\sum_s q_k z_k x_k^{'} \right)^{-1} \sum_s q_k z_k y_k \tag{2.49}$$

Capítulo 3

Variables instrumentales

3.1. Metodología

Segun (Angrist & Pischke (2008) y Bernal & Peña (2010)), este método busca eliminar el sesgo de selección ocasionado por la no aleatoriedad en la selección del tratamiento, el cual genera que exista correlación entre la variable indicadora de tratamiento T_i y otras variables no observables o el termino de error ε_i. Por lo tanto se define una variable Z_i denominado instrumento que cumple con las siguientes condiciones:

$$(\textit{Relevancia del instrumento }) Cov(T_i, Z_i) \neq 0$$

$$(\textit{Exogeneidad del instrumento }) Cov(\varepsilon_i, Z_i) = 0$$

De modo que se tiene un subconjunto de individuos los cuales determinan su participación en el programa T_i debido a la variable instrumental Z_i. Por lo cual la estimación del efecto no representara a todos los tratados, es decir que se estima un efecto promedio local ($LATE$). Sin embargo esta salvedad es relevante en el caso de que los efectos sean heterogéneos. Puesto que en dado caso es necesario que la variable instrumental cumpla con la condición de (*Monotonicidad*), la cual considera que la variable indicadora de tratamiento T_i es una función monótona creciente de Z_i. En caso contrario el efecto local será igual al (ATE).

Considerando las anteriores condiciones y teniendo en cuenta la expresión de la variable resultado en la ecuación (2.1), se realiza una estimación por el método de mínimos cuadrados en dos etapas o por máxima verosimilitud dos veces. Entonces se procedería de la siguiente manera:

Se realiza una regresión de la variable T_i sobre la variable instrumental Z_i estimándose por alguna de las metodologías mencionadas, construyendo de este modo el estimador del indicador de participación en el programa $\hat{T}_i$

$$\hat{T}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_i$$

Seguidamente se procede a estimar la regresión (2.1) sustituyendo T_i por $\hat{T}_i$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{T}_i$$

Obteniendo finalmente un estimador del impacto del programa $\hat{\beta}_1$, insesgado y consistente.

3.1.1. Estimación

Suponga un vector de N variables aleatorias $Y = (Y_1, Y_2, \cdots, Y_N)^{'}$, y otro vector de variables aleatorias $\mathbf{T}_k = (\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_N)^{'}$, los cuales se relacionan por el siguiente modelo de probabilidad ξ

$$Y_k = \beta_0 + \beta_1 \mathbf{T}_k + \varepsilon_k \tag{3.1}$$

Donde el vector de variables aleatorias $\mathbf{T}_k$ indica la participación en el programa. Sin embargo este presenta problemas de endogeneidad, por lo tanto se considera la existencia de un vector de variables instrumentales aleatorias $Z = (Z_1, Z_2, \cdots, Z_N)$, la cual cumple con las condiciones mencionadas anteriormente y se relaciona con la variable $\mathbf{T}_k$ por un modelo:

$$\mathbf{T}_k = \gamma_0 + \gamma_1 Z_k + \omega_k \tag{3.2}$$

De modo que los valores estimados de $\mathbf{T}_k$ obtenidos a partir de este ya no presentan endogeneidad y se reemplazan en la estimación de la ecuación (3,1).

Dadas las anteriores salvedades, para la ecuación (3.1) se supone que cada ε_k con $k \in U$, son variables aleatorias independientes e idénticamente distribuidos con media cero y varianza $c_k\sigma^2$, bajo las cuales se tienen las siguientes propiedades:

$$E_\xi(Y_k) = \mathbf{T}_k^{'}\beta \tag{3.3}$$

$$Var_\xi(Y_k) = c_k\sigma^2 \tag{3.4}$$

Análogamente se tiene las mismas propiedades para la ecuación (3.2).

Considerando lo mencionado se tienen las variables aleatorias $y = (y_1, y_2, \cdots, y_N)$ y $T = (T_1, T_2, \cdots, T_N)$ que serán realizaciones de los vectores aleatorios Y_k y $\mathbf{T}_k$, para los que se tienen los siguientes modelos en la población finita N:

$$y_k = B_0 + B_1T_k + E_k \tag{3.5}$$

$$T_k = G_0 + G_1z_k + U_k \tag{3.6}$$

Teniendo en cuenta que, para estimar los coeficientes de estas regresiones se tendría que realizar un censo y en la práctica esto no es factible se debe tomar una muestra de las realizaciones de las variables aleatorias que la integran. Por lo tanto se obtienen los siguientes modelos para estas:

$$y_k = \hat{B}_0 + \hat{B}_1\hat{T}_k + e_k \tag{3.7}$$

$$\hat{T}_k = \hat{G}_0 + \hat{G}_1z_k + u_k \tag{3.8}$$

A partir de los cuales se logra estimar B_1 por medio de la estimación de $\hat{B}_1$, utilizando la metodología de variables instrumentales.

Como se planteó anteriormente se toma una muestra de tamaño n la cual, es seleccionada bajo un diseño muestral con una distribución de probabilidad multivariante $p(s) > 0$ definida para un soporte Q.

Bajo este diseño muestral $p(\cdot)$, se asignan probabilidades de inclusión de primer y segundo orden dados por

$$\pi_k = P(k \in s) = \sum_{s \ni k} p(s) \tag{3.9}$$

$$\pi_{kl} = P(k \in syl \in s) = \sum_{s \ni kyl} p(s) \tag{3.10}$$

Donde $s \ni k$ hace referencia a la suma en todas las muestras que contienen el k-ésimo elemento. Del mismo modo para π_{kl} se tendrá la probabilidad de que k y l pertenezcan a una misma muestra. También se tiene $d_k = \frac{1}{\pi_k}$ que es el inverso de la probabilidad de inclusión.

Ahora bien, dadas estas pautas se procede con la estimación de la siguiente manera:

Primera etapa :

Se realiza la estimación para el modelo de la ecuación (3.8), para el cual se supone que la relación no pasa por el origen y se tiene una estructura de varianza constante para todo $k \in U$. Dado esto se tiene que $c_k = 1$ y el termino de error se asume como variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ_T^2.
Finalmente se tiene que la estimación de $\hat{T}_k$ estará dada por:

$$\hat{\tilde{T}}_k = \hat{\tilde{G}}_0 + \hat{\tilde{G}}_1 z_k \tag{3.11}$$

y sus coeficientes se obtienen por las siguientes expresiones:

$$\hat{\tilde{G}}_1 = \frac{\sum_s \frac{(z_k - \tilde{z_s})(\hat{T}_k - \tilde{T_s})}{\pi_k}}{\sum_s \frac{(z_k - \tilde{z_s})^2}{\pi_k}} \tag{3.12}$$

$$\hat{\tilde{G}}_0 = \tilde{T_s} - \hat{\tilde{G}}_1 \tilde{z_s} \tag{3.13}$$

Donde

$$\tilde{z_s} = \frac{\sum_s \frac{z_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}}$$

$$\tilde{T}_s = \frac{\sum_s \frac{T_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}}$$

Son los estimadores de Hájek para la media. Esta estimación será adecuada cuando el indicador de tratamiento en el programa T_k toma valores continuos, si es dicotómica puede ser estimado por un modelo logit o probit, en el primer caso se tendría el siguiente modelo:

$$\hat{T_k} = \frac{1}{1 + e^{-(\hat{G_0} + \hat{G_1} z_{1k})}} + u_k \tag{3.14}$$

El cual al ser no lineal tiene que ser estimado por máxima verosimilitud que en poblaciones finitas se utiliza la máxima pseudo-verosimilitud. Entonces se tiene que al ser $\hat{T} \sim bernoulli(p)$ donde el parámetro p toma la siguiente forma:

$$p \equiv Pr(\hat{T} = 1 \mid z) = F(z_k^{'}\hat{G}) = \frac{e^{(\hat{G_0} + \hat{G_1} z_{1k})}}{1 + e^{(\hat{G_0} + \hat{G_1} z_{1k})}} \tag{3.15}$$

por lo tanto la función de pseudo-verosimilitud es:

$$L(\hat{G}, \hat{T}) = \prod_s (p_k^{d_k \hat{T_k}})(1 - p_k)^{d_k(1-\hat{T_k})} \tag{3.16}$$

de manera que:

$$l(\hat{G}, \hat{T}) = \sum_s d_k [\hat{T_k} ln F(z_k^{'}\hat{G}) + (1 - \hat{T_k}) ln(1 - F(z_k^{'}\hat{G}))] \tag{3.17}$$

al derivar con respecto a $\hat{G}$ se obtienen q sistemas de ecuaciones no-lineales con q incógnitas que se solucionan por métodos numéricos.

Segunda etapa :

En esta etapa con los valores obtenidos anteriormente se realiza la estimación de la ecuación (3.7), para el cual suponemos una relación que no pasa por el origen y una varianza constante para todo $k \in U$, con un término de error aleatorio, independiente e idénticamente distribuido que tiene media cero y varianza σ^2.
Entonces la estimación por MCO será:

$$\hat{y_k} = \hat{\hat{B}}_0 + \hat{\hat{B}}_1 \hat{\hat{T}}_k \tag{3.18}$$

Para la cual los coeficientes $\hat{\hat{B}}_i$ están dados por:

$$\hat{\hat{B}}_1 = \frac{\sum_s d_k(\hat{\hat{T}}_k - \tilde{\hat{\hat{T}}}_s)(y_k - \tilde{y_s})}{\sum_s d_k(\hat{\hat{T}}_k - \tilde{\hat{\hat{T}}}_s)^2} \tag{3.19}$$

$$\hat{\hat{B}}_0 = \tilde{y_s} - \hat{\hat{B}}_1 \tilde{\hat{\hat{T}}}_s \tag{3.20}$$

donde $\tilde{y_s}$ y $\tilde{\hat{\hat{T}}}_s$ son los estimadores de Hájek para la media que están dados por las siguientes expresiones :

$$\tilde{y_s} = \frac{\sum_s d_k y_k}{\sum_s d_k}$$

$$\tilde{\hat{\hat{T}}}_s = \frac{\sum_s d_k \hat{\hat{T}}_k}{\sum_s d_k}$$

Expresándolo de modo matricial tendremos que:

$$\hat{\hat{B}}_{reg} = \left[\boldsymbol{T}' \boldsymbol{P_z} \boldsymbol{d_k} \boldsymbol{T}\right]^{-1} \boldsymbol{T}' \boldsymbol{P_z} \boldsymbol{d_k} \boldsymbol{Y} \tag{3.21}$$

Dónde:

La matriz $\boldsymbol{T}$ en este caso contiene un vector de unos y la variable indicadora de tratamiento.

$$\boldsymbol{T} = \begin{bmatrix} 1 & T_{1k} \\ \vdots & \vdots \\ 1 & T_{nk} \end{bmatrix} \tag{3.22}$$

$$\boldsymbol{P_z} = \boldsymbol{Z}\left[\boldsymbol{Z}'\boldsymbol{Z}\right]^{-1}\boldsymbol{Z}' \tag{3.23}$$

para la cual $\boldsymbol{Z}$ representa una matriz conformada con las variables instrumentales de las respectivas variables explicativas.

$$\boldsymbol{Z} = \begin{bmatrix} 1 & Z_{1k} \\ \vdots & \vdots \\ 1 & Z_{nk} \end{bmatrix} \tag{3.24}$$

$\boldsymbol{d_k}$ es una matriz diagonal del inverso de las probabilidades de inclusión.

$$\boldsymbol{d_k} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix} \tag{3.25}$$

Finalmente tenemos el vector $\boldsymbol{Y}$ que contiene los respectivos valores de la variable de interés.

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{3.26}$$

Cabe resaltar que las anteriores expresiones son válidas en el caso de que la variable indicadora de tratamiento no sea dicotómica, ya que de ser así se obtendría un modelo lineal de probabilidad (MLP), que como se sabe incide en los siguientes problemas:

- Violación del supuesto de homocedasticidad.
- Obtención de predicciones de probabilidad mayores que uno y menores que cero.

Por lo tanto el estimador estará dado por:

$$\hat{\hat{B}}_{reg} = \left[\hat{\boldsymbol{T}}' \boldsymbol{d_k} \hat{\boldsymbol{T}}\right]^{-1} \hat{\boldsymbol{T}}' \boldsymbol{d_k} \boldsymbol{Y} \tag{3.27}$$

Donde $\hat{\boldsymbol{T}}$ es una matriz que contiene el vector de unos y el vector de predicciones de las probabilidades obtenidas del modelo logit o probit.

Capítulo 4

Estimación con pesos calibrados

Ahora consideramos reemplazar los pesos d_k por ponderadores más eficientes como son los pesos calibrados w_k que están determinados por la información auxiliar disponible y cumplen con las condiciones mencionadas anteriormente. Dado esto tenemos que:

$$\sum_{k \in s} w_k \hat{\hat{T}}_k = t_{\hat{\hat{T}}} \tag{4.1}$$

Este sistema de ponderadores puede ser construido bajo el enfoque de minimización de la distancia o el enfoque funcional. En este caso se usara el segundo, para el cual los ponderadores de calibración son definidos como:

$$w_k = d_k(1 + \lambda^{'} z_k) \tag{4.2}$$

donde $q_k = d_k$ que son los pesos dados por el diseño muestral, z_k es el vector de variables instrumentales conocido para todo $k \in s$ y el vector λ está dado por

$$\lambda^{'} = (t_{\hat{\hat{T}}} - \hat{t}_{\hat{\hat{T}}\pi})^{'} (\sum_{k \in s} d_k z_k \hat{\hat{T}}_k^{'})^{-1} \tag{4.3}$$

De donde se tiene que $t_{\hat{\hat{T}}}$ y $\hat{t}_{\hat{\hat{T}}\pi}$ son el total conocido y el total estimado por Horvitz - Thompson, de $\hat{\hat{T}}$ respectivamente.

Entonces aplicando los ponderadores w_k tendremos que los coeficientes de regresión estarán dados por las siguientes expresiones:

$$\hat{\hat{B}}_{calf} = \left[\boldsymbol{T}'\boldsymbol{P_z}\boldsymbol{w_k}\boldsymbol{T}\right]^{-1}\boldsymbol{T}'\boldsymbol{P_z}\boldsymbol{w_k}\boldsymbol{Y} \tag{4.4}$$

Con $\boldsymbol{w_k}$ que será la matriz diagonal que contiene los pesos de calibración y estará dado por

$$\boldsymbol{w_k} = \begin{bmatrix} w_{11} & 0 & \cdots & 0 \\ 0 & w_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{nn} \end{bmatrix} \tag{4.5}$$

En el caso de que la variable indicadora de tratamiento sea dicotómica, se procederá análogamente como se explicó en el capítulo anterior aplicando los pesos de calibración por enfoque funcional. Obteniendo de este modo para los coeficientes la siguiente expresión:

$$\hat{\hat{B}}_{calf} = \left[\hat{\boldsymbol{T}}'\boldsymbol{w_k}\hat{\boldsymbol{T}}\right]^{-1}\hat{\boldsymbol{T}}'\boldsymbol{w_k}\boldsymbol{Y} \tag{4.6}$$

4.0.1. Modelo con otras covariables X

En la práctica pueden presentarse algunas otras covariables auxiliares que permiten describir mejor el comportamiento de y_k entonces tendremos que

$$\boldsymbol{Y} \quad = B_0 + B_1\hat{T}_k + X_kB_{ad} + E_{k,ext} \tag{4.7}$$

Donde X_k es una matriz de covariables de las cuales conocemos sus totales, además se tiene respectivamente una matriz de variables instrumentales con igual dimensión. Por lo tanto reescribiendo (4,7) tenemos que

$$\boldsymbol{Y} = \boldsymbol{T}'_{k,ext}\boldsymbol{B}_{ext} + \boldsymbol{E}_{k,ext} \tag{4.8}$$

De manera que $\boldsymbol{T}_{k,ext}$ es una matriz compuesta por un vector de unos, el estimador de la variable indicadora de tratamiento y las covariables x_k, con el vector $\boldsymbol{B}_{ext}$ que contiene los respectivos coeficientes y $\boldsymbol{E}_{k,ext}$ será el

vector de residuales de este modelo.

Con lo cual tendremos que la estimación estará dada por

$$\hat{\boldsymbol{Y}} = \boldsymbol{T}_{k,ext}^{'}\hat{\boldsymbol{B}}_{ext} \tag{4.9}$$

donde

$$\hat{\boldsymbol{B}}_{ext} = \left(\boldsymbol{T}_{k,ext}^{'}\boldsymbol{w}_k\boldsymbol{T}_{k,ext}\right)^{-1}\boldsymbol{T}_{k,ext}^{'}\boldsymbol{w_k}\boldsymbol{Y} \tag{4.10}$$

con $\boldsymbol{w}_k$ la matriz de ponderadores calibrados que para cada individuo k están dados por

$$w_k = d_k\left[1 + \left(t_{\hat{\hat{T}}} - \hat{t}_{\hat{\hat{T}}\pi}\right)^{'}\left(\sum_{k\in s} d_k z_k T_{k,ext}^{'}\right)^{-1} z_k\right] \tag{4.11}$$

donde $t_{\hat{\hat{T}}}$ es el vector de totales conocidos para cada covariable, $\hat{t}_{\hat{\hat{T}}\pi}$ es el vector de totales estimados por Horvitz-Thompson, los z_k representan las variables instrumentales respectivas y $T_{k,ext}$ serán los valores de las covariables.

Sin embargo en la práctica no se suele conocer el vector de totales y un conjunto de variables instrumentales para otras covariables, por lo tanto no se pueden implementar en la calibración, aun así podemos incluirlas en la regresión logrando de este modo explicar una mayor variación de la variable respuesta que la que se explicaría solo usando la covariable de calibración. Pero esto implica que la esperanza poblacional de los residuales para cada k sea distinta de cero.

4.1. Inferencia sobre los parametros

Resulta de interés conocer la significación estadística en el modelo de los parámetros $\hat{B}$ para lo cual se realizan intervalos de confianza y pruebas de hipótesis. Por ende se requiere saber si son insesgados, conocer la varianza y distribución de probabilidad de estos.
Por lo tanto bajo los supuestos del término de error $E(e_i) = 0$, $V(e_i) = \sigma^2$ $\forall\ i = 1, \cdots, n$ y $cov(e_i, e_j) = 0$, para $i \neq j$, se comprueba si el parámetro

es insesgado:
Reescribimos el coeficiente $\hat{\hat{B}}$ de la siguiente manera

$$\hat{\hat{B}}_{calf,MC2E} = \hat{\boldsymbol{T}}^{-1}_{calf,MC2E}\hat{\boldsymbol{t}}_{calf,MC2E} \tag{4.12}$$

donde

$$\hat{T}_{calf,MC2E} = \boldsymbol{T}'\boldsymbol{P_z}\boldsymbol{w_k}\boldsymbol{T} \tag{4.13}$$

$$\hat{t}_{calf,MC2E} = \boldsymbol{T}'\boldsymbol{P_z}\boldsymbol{w_k}\boldsymbol{Y} \tag{4.14}$$

para las anteriores expresiones se tiene que cada elemento que compone estas matrices, estará respectivamente dado por

$$\hat{t}_{jj',calf} = \sum_{k\in s} T_{jk}P_{z,jk}w_kT_{j'k} \tag{4.15}$$

$$\hat{t}_{j0,calf} = \sum_{k\in s} T_{j'k}P_{z,jk}w_ky_k \tag{4.16}$$

Dada la ecuación (4,7) consideramos la esperanza de esta para verificar la insesgades de los estimadores $\hat{\hat{B}}$

$$E($$

4.2. Estimación aplicando el diseño muestral

4.2.1. Diseño de muestreo aleatorio simple

Este es un diseño simple para el cual los pesos muestrales $d_k = \frac{N}{n}$ donde los pesos de calibración están dados por

$$w_k = \frac{N}{n}\left(1 + \lambda' z_k\right) \tag{4.18}$$

$$\lambda = \left(t_{\hat{\hat{T}}} - \frac{N}{n}\sum_{k\in s}\hat{\hat{T}}_k\right)\left(\frac{N}{n}\sum_{k\in s} z_k\hat{\hat{T}}'_k\right)^{-1} \tag{4.19}$$

$$\hat{\hat{B}}_{calf} = \left(\hat{\boldsymbol{T}}_{calf,MC2E}\right)^{-1}\hat{\boldsymbol{t}}_{calf,MC2E}$$

$$AV\left(\hat{\hat{B}}_{calf}\right) = \boldsymbol{T}^{-1}\boldsymbol{V}\boldsymbol{T}^{-1}$$

Luego los elementos v_{jj} que componen la matriz de varianzas y covarianzas están dados por

$$v_{jj} = N^2\left(1 - \frac{n}{N}\right)S^2_{H_k}n^{-1} \tag{4.20}$$

donde $S^2_{H_k}$ y $\bar{H}_U$ son respectivamente la varianza y media poblacional de H_k dada por

$$S^2_{H_k} = (N-1)^{-1}\sum_U \left(H_k - \bar{H}_U\right)^2$$

$$\bar{H}_U = N^{-1}\sum_U H_k$$

$$H_k = T_{jk}P_{zjk}E_k - T_{j'k}Q \tag{4.21}$$

Un estimador de esta aproximación de varianza está dado por

$$\hat{V}\left(\hat{\hat{B}}_{calf}\right) = \hat{\boldsymbol{T}}^{-1}_{calf,MC2E}\hat{\boldsymbol{V}}\hat{\boldsymbol{T}}^{-1}_{calf,MC2E}$$

los elementos $\hat{v}_{jj}$ que componen la matriz $\hat{\boldsymbol{V}}$ están dados por

$$\hat{v}_{jj} = N^2\left(1 - \frac{n}{N}\right)S^2_{h_k}n^{-1} \tag{4.22}$$

de donde

$$S^2_{h_k} = (N-1)^{-1}\sum_s \left(h_k - \bar{h}_s\right)^2$$

$$\bar{h}_s = N^{-1}\sum_s h_k$$

con

$$h_k = T_{jk}P_{zjk}e_k - T_{j'k}\hat{Q} \tag{4.23}$$

donde e_k son los residuos muestrales

$$e_k = y_k - T_k^{'}\hat{\hat{B}}_{calf}$$

y $\hat{Q}$ viene dado por

$$\hat{Q} = \left(\frac{N}{n}\sum_s z_k T_{j'k}\right)^{-1}\frac{N}{n}\sum_s z_k T_{jk}P_{zjk}e_k$$

4.2.2. Diseño Estratificado MAS

Para un diseño de muestreo aleatorio estratificado, se tiene para cada k que pertenece al estrato h el ponderador $d_k = \frac{N_h}{n_h}$, por ende los ponderadores de calibración se definen como

$$w_k = \frac{N_h}{n_h}\left(1 + \lambda^{`} z_k\right) \tag{4.24}$$

Considerando lo anterior el estimador de los coeficientes será

$$\hat{\hat{B}}_{calf} = \left(\hat{\boldsymbol{T}}_{calf,MC2E}\right)^{-1} \hat{\boldsymbol{t}}_{calf,MC2E}$$

De donde se tiene que cada elemento de las respectivas matrices está dado por

$$\hat{t}_{jj',calf} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k\in s} T_{jk} P_{z,jk}\left(1 + \lambda^{`} z_k\right) T_{j'k} \tag{4.25}$$

$$\hat{t}_{j0,calf} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k\in s} T_{j'k} P_{z,jk}\left(1 + \lambda^{`} z_k\right) y_k \tag{4.26}$$

luego el vector λ se definen como

$$\lambda_t^{`} = \left(t_{\hat{\hat{T}}} - \hat{t}_{\hat{\hat{T}}_\pi}\right)^{`} \left(\frac{N_h}{n_h} \sum_{k\in s} z_k \hat{\hat{T}}_k^{`}\right)^{-1} \tag{4.27}$$

Con $\hat{t}_{\hat{\hat{T}}_\pi}$ hace referencia a el estimador Horvitz – Thompson para un muestreo estratificado simple.

Se da una aproximación de la varianza para el estimador bajo este diseño

$$AV\left(\hat{\hat{B}}_{calf}\right) = \boldsymbol{T}^{-1}\boldsymbol{V}\boldsymbol{T}^{-1}$$

Donde los elementos de la matriz V se definen como

$$v_{jj} = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{N_h}{n_h}\right) S^2_{H_{k,h}} n_h^{-1} \tag{4.28}$$

de donde $S^2_{H_{k,h}}$ es la varianza de H_k para cada estrato

$$S^2_{H_{k,h}} = (N_h – 1)^{-1} \sum_{U_h} \left(H_k - \bar{H}_{U_h}\right)^2 \tag{4.29}$$

con $H_k = T_{jk} P_{zjk} E_k - T_{j'k} Q$ y su media poblacional $\bar{H}_{U_h} = N_h^{-1} \sum_{U_h} H_k$

Se define un estimador de la varianza

$$\hat{V}\left(\hat{\hat{B}}_{calf}\right) = \hat{\boldsymbol{T}}^{-1}_{calf,MC2E} \hat{\boldsymbol{V}} \hat{\boldsymbol{T}}^{-1}_{calf,MC2E}$$

donde los elementos $\hat{v}_{jj}$ de la matriz $\hat{\boldsymbol{V}}$ están dados por

$$\hat{v}_{jj} = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) S^2_{h_{k,h}} n_h^{-1} \tag{4.30}$$

con

$$S^2_{h_{k,h}} = (N_h – 1)^{-1} \sum_{s_h} \left(h_k - \bar{h}_{s_h}\right)^2 \tag{4.31}$$

$$h_k = T_{jk} P_{zjk} e_k - T_{j'k} \left(\frac{N_h}{n_h} \sum_{s_h} z_k T_{j'k}\right)^{-1} \frac{N_h}{n_h} \sum_{s} z_k T_{jk} P_{zjk} e_k \tag{4.32}$$

donde e_k son los residuos muestrales

$$e_k = y_k - T_k^{'} \hat{\hat{B}}_{calf}$$

$$e_k = y_k - \hat{\hat{T}}_k^{`} \left(\sum_{k \in s_h} \frac{N_h}{n_h} z_k \hat{\hat{T}}_k^{`}\right)^{-1} \left(\sum_{k \in s_h} \frac{N_h}{n_h} z_k y_k\right)$$

4.2.3. Diseño bietápico MAS-MAS

Bajo el cumplimiento de las propiedades de invarianza y de independencia se formula un estimador en dos etapas que aplica para las unidades primarias un diseño muestral MAS y para la segunda etapa estima el total de la variable de interés por calibración por enfoque funcional bajo un diseño MAS.
Por lo tanto el estimador estará dado de la siguiente manera:

$$\hat{\hat{B}}_{calf} = \left(\hat{\boldsymbol{T}}_{calf,MC2E}\right)^{-1} \hat{\boldsymbol{t}}_{calf,MC2E}$$

Con cada elemento que compone las matrices $\hat{\boldsymbol{T}}_{calf,MC2E}$ y $\hat{\boldsymbol{t}}_{calf,MC2E}$ definido respectivamente por

$$\hat{t}_{jj',calf} = \frac{N_I}{n_I} \sum_{I \in S_I} \frac{N_i}{n_i} \sum_{k \in s_i} T_{jk} P_{z,jk} \left(1 + \lambda^{`} z_k\right) T_{j'k} \tag{4.33}$$

$$\hat{t}_{j0,calf} = \frac{N_I}{n_I} \sum_{I \in S_I} \frac{N_i}{n_i} \sum_{k \in s_i} T_{j'k} P_{z,jk} \left(1 + \lambda^{`} z_k\right) y_k \tag{4.34}$$

donde $\lambda^{`} = \left(t_{\hat{\hat{T}}} - \frac{N_i}{n_i} \sum_{k \in s_i} \hat{\hat{T}}_k\right)^{`} \left(\frac{N_i}{n_i} \sum_{k \in s_i} z_k \hat{\hat{T}}_k^{`}\right)^{-1}$

Se define una aproximación de la varianza

$$AV\left(\hat{\hat{B}}_{calf}\right) = \boldsymbol{T}^{-1} \boldsymbol{V} \boldsymbol{T}^{-1}$$

donde los elementos que componen la matriz $\boldsymbol{V}$ están dados por

$$v_{jj} = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S^2_{\hat{\hat{B}}_{calf},U_I} + \frac{N_I}{n_I} \sum_{I \in U_I} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S^2_{H_k,U_i} \tag{4.35}$$

cuya estimación de la varianza es

$$\hat{V}\left(\hat{\hat{B}}_{calf}\right) = \hat{\boldsymbol{T}}^{-1}_{calf,MC2E} \hat{\boldsymbol{V}} \hat{\boldsymbol{T}}^{-1}_{calf,MC2E}$$

con elementos $\hat{v}_{jj}$ de la matriz $\hat{\boldsymbol{V}}$ dados por

$$\hat{v}_{jj} = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S^2_{\hat{\hat{B}}_{calf},S_I} + \frac{N_I}{n_I} \sum_{I \in S_I} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S^2_{h_k,s_i} \tag{4.36}$$

donde $S^2_{\hat{\hat{B}}_{calf},U_I}$ y $S^2_{H_k,U_i}$ hacen referencia, a la varianza poblacional en todas las unidades primarias para los coeficientes y la varianza poblacional entre los H_k dentro de cada unidad primaria de muestreo respectivamente. Similarmente se tiene en la muestra $S^2_{\hat{\hat{B}}_{calf},S_I}$ y $S^2_{h_k,s_i}$.

Capítulo 5

Simulación

Con el objetivo de conocer que tan eficiente es el estimador propuesto se observa el sesgo relativo y la eficiencia relativa, por lo tanto se ha realizado simulaciones de Monte Carlo, considerando para la comparación el estimador de regresión Horvitz-Thompson. Por ende se genera una población finita de tamaño $N = 10000$ por el modelo de superpoblación

$$Y_k = \beta_0 + \beta_1 T_k + \varepsilon_k$$

De donde consideramos la existencia de sesgo por autoselección en el programa, lo cual se presenta por omisión de variables en la regresión. También se genera una variable instrumental Z_k y se asume como conocido el total poblacional para la variable explicativa.

5.0.1. Especificación de las poblaciones y muestras

Considerando las afirmaciones anteriores las poblaciones se simularon bajo los siguientes parámetros: Se plantearon poblaciones con una variable indicadora de tratamiento T_k dicotómica y con una variable T_k continua. Para las cuales se utilizaron variables instrumentales Z_k que cumplieran con la condición de exogeneidad del instrumento, además se plantea escenarios donde se cumplen con la condición de relevancia y otros en los que no.

Partiendo de los escenarios anteriores se realizaron ejercicios de Monte Carlo de tamaño $nsim = 1000$. Para el primer escenario donde la variable T_k

es dicotómica, las estimaciones y selección de la muestra se realizan bajo los diseños muestrales MAS y estratificado MAS. Para el ultimo escenario, donde la variable T_k es continua, las estimaciones y selección de la muestra se hacen bajo los diseños MAS, estratificado MAS y πPT.

Para los distintos escenarios se realizan las simulaciones en el programa estadístico R, repitiendo los siguientes procesos mil veces:

- Se selecciona según el escenario una muestra aleatoria bajo diseño MAS, estratificado MAS o πPT, de la población simulada.
- Según el escenario se realiza una regresión sobre T_k con Z_k como variable explicativa.
- Se obtienen los pesos de calibración w_k bajo el enfoque funcional y se obtienen los coeficientes $\hat{\hat{B}}_{calf}$.
- Finalmente se guardan las estimaciones y se calculan las siguientes medidas.

  a promedio de los coeficientes simulados: definidos como

$$mean(\hat{\hat{B}}) = \frac{1}{nsim} \sum_{i=1}^{nsim} \hat{\hat{B}} \tag{5.1}$$

  b Sesgo relativo: definido como

$$BR(\hat{\hat{B}}_{calf}) = \frac{1}{nsim} \sum_{i=1}^{nsim} \left(\frac{\hat{\hat{B}}_{calf} - B}{B} \right) \tag{5.2}$$

  c Eficiencia relativa: definida como

$$ER(\hat{\hat{B}}_{calf}) = \frac{\sum_{i=1}^{nsim} (\hat{\hat{B}}_{calf} - B)^2}{\sum_{i=1}^{nsim} (\hat{\hat{B}}_{\pi} - B)^2} \tag{5.3}$$

5.0.2. Resultados de las simulaciones

En el escenario que se consideró la variable indicadora de tratamiento T_k dicotómica, donde la variable instrumental cumple con las condiciones de relevancia y exogeneidad. Se encontró que el estimador propuesto de calibración bajo un diseño MAS presenta estimaciones promedio muy similares al estimador clásico, además se logra notar sesgos relativos menores y es más eficiente que el estimador por Horvitz-Thompson. En el cuadro (5,1) notamos que para muestras de tamaño 50, el estimador propuesto con una primera etapa por MLP presenta estimaciones con poco sesgo y eficientes siendo de este modo recomendable aplicar el estimador de calibración bajo dichas condiciones.

Cuadro 5.1: Simulación $N = 10000$, $n = 50$

		Promedio de los coeficientes			
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	2,018876	2,041311	1,98399	1,925365
	$\hat{\hat{B}}_1$	6,966437	6,933906	7,009669	7,087809

		Sesgo relativo				Eficiencia relativa	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	1,018876	1,041311	0,9839896	0,9253649	0,7194485	3,264009
	$\hat{\hat{B}}_1$	5,966437	5,933906	6,009669	6,087809	0,08298796	1,765044

Cuadro 5.2: Simulación $N = 10000$, $n = 100$

		Promedio de los coeficientes			
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	1,994192	2,01925	1,979606	1,952026
	$\hat{\hat{B}}_1$	6,999464	6,962987	7,01905	7,057565

		Sesgo relativo				Eficiencia relativa	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	0,9941922	1,01925	0,9796062	0,9520259	12,33006	6,210998
	$\hat{\hat{B}}_1$	5,999464	5,962987	6,01905	6,057565	1265,054	2,418848

Cuando se tienen tamaños de muestra $n = 1000$ (Ver cuadro 5,3), el estimador de calibración con primera etapa por modelo logit presenta sesgos relativos pequeños y eficientes incluso cuando la variable instrumental resulta irrelevante mostrando que es recomendable el uso del estimador de calibración.

Cuadro 5.3: Simulación $N = 10000$, $n = 1000$

		Promedio de los coeficientes			
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	1,997024	2,021409	1,994317	1,999157
	$\hat{\hat{B}}_1$	6,994141	6,960449	6,998037	6,991144
Z_k irrelevante	$\hat{\hat{B}}_0$	1,826965	1,83631	1,237545	1,858191
	$\hat{\hat{B}}_1$	7,286523	7,26961	8,367139	7,274533

		Sesgo relativo				Eficiencia relativa	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	0,9970237	1,021409	0,9943171	0,9991574	3,645804	0,001549207
	$\hat{\hat{B}}_1$	5,994141	5,960449	5,998037	5,991144	0,1122991	0,05014013
Z_k irrelevante	$\hat{\hat{B}}_0$	0,8269652	0,8363098	0,2375445	0,8581906	19,4161	0,7505242
	$\hat{\hat{B}}_1$	6,286523	6,26961	7,367139	6,274533	22,76709	1,036854

En los cuadros (5,4) y (5,5) se muestran los resultados bajo un diseño estratificado MAS con tamaños muestrales 200 y 2000. Para estos escenarios encontramos que el estimador propuesto no presenta tan buenos estimadores y es menos eficiente que el estimador normalmente aplicado.

Cuadro 5.4: Simulación $N = 10000$, $n = 200$

		Promedio de los coeficientes			
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	1,988939	2,010702	2,295211	2,53191
	$\hat{\hat{B}}_1$	7,014169	6,984197	6,715522	6,389919
Z_k irrelevante	$\hat{\hat{B}}_0$	1,338667	1,347925	1,165883	−107,4851
	$\hat{\hat{B}}_1$	8,202427	8,18538	8,534709	216,1979

		Sesgo relativo				Eficiencia relativa	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	0,09889389	0,1010702	0,1295211	0,153191	712,3068	2470,395
	$\hat{\hat{B}}_1$	0,6014169	0,5984197	0,5715522	0,5389919	1,590794	28191,29
Z_k irrelevante	$\hat{\hat{B}}_0$	0,338667	0,3479251	0,1658827	−108,4851	712,3068	2470,395
	$\hat{\hat{B}}_1$	7,202427	7,18538	7,534709	215,1979	1,629051	31145,78

Cuadro 5.5: Simulación $N = 10000$, $n = 2000$

		Promedio de los coeficientes			
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	1,978909	1,984065	2,257351	2,431748
	$\hat{\hat{B}}_1$	7,026237	7,019136	6,759364	6,509488
Z_k irrelevante	$\hat{\hat{B}}_0$	−0,1448881	−0,133799	1,15012	115,1977
	$\hat{\hat{B}}_1$	10,9451	10,92463	8,561893	−191,2996

		Sesgo relativo				Eficiencia relativa	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{\pi,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,PVMCO}$
Z_k relevante	$\hat{\hat{B}}_0$	0,9789094	0,9840651	1,257351	1,431748	712,3068	2470,395
	$\hat{\hat{B}}_1$	6,026237	6,019136	5,759364	0,5389919	403,1061	1490,323
Z_k irrelevante	$\hat{\hat{B}}_0$	−1,144888	−1,133799	0,1501195	114,1977	0,1570025	2814,287
	$\hat{\hat{B}}_1$	9,945095	9,924632	7,561893	−192,2996	0,1567428	2552,969

Consideramos ahora el escenario con variable indicadora T_k continua, donde la variable instrumental cumple la condición de exogeneidad bajo un diseño MAS con tamaños muestrales 50 y 100 (Ver cuadros 5,6 y 5,7) para estos casos el estimador propuesto presenta resultados similares a los del estimador Horvitz-Thompson, tanto para el caso de variable instrumental relevante e irrelevante. En el cuadro (5,8) para tamaños muestrales $n = 1000$ se tiene que el estimador IV calibrado presenta poco sesgo relativo y es más eficiente que el estimador común.

Cuadro 5.6: Simulación $N = 10000$, $n = 50$

		Promedio de los coeficientes	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	2,002997	2,003838
	$\hat{\hat{B}}_1$	7,000017	6,999936
Z_k irrelevante	$\hat{\hat{B}}_0$	1,48012	0,505506
	$\hat{\hat{B}}_1$	7,186257	7,920505

		Sesgo relativo		Eficiencia relativa
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	1,002997	1,003838	1,639906
	$\hat{\hat{B}}_1$	6,000017	5,999936	14,40069
Z_k irrelevante	$\hat{\hat{B}}_0$	0,4801201	−0,494494	8,263849
	$\hat{\hat{B}}_1$	6,186257	6,920505	24,42468

Cuadro 5.7: Simulación $N = 10000$, $n = 100$

		Promedio de los coeficientes	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	2,006497	2,006629
	$\hat{\hat{B}}_1$	7,00008	7,000117
Z_k irrelevante	$\hat{\hat{B}}_0$	1,599373	1,385838
	$\hat{\hat{B}}_1$	7,176261	7,256802

		Sesgo relativo		Eficiencia relativa
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	1,006497	1,006629	1,040952
	$\hat{\hat{B}}_1$	6,00008	6,000117	2,144172
Z_k irrelevante	$\hat{\hat{B}}_0$	0,599373	0,3858382	2,350093
	$\hat{\hat{B}}_1$	6,176261	6,256802	2,122694

Cuadro 5.8: Simulación $N = 10000$, $n = 1000$

		Promedio de los coeficientes	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	2,001664	2,00161
	$\hat{\hat{B}}_1$	7,000276	7,000277
Z_k irrelevante	$\hat{\hat{B}}_0$	1,454747	1,595413
	$\hat{\hat{B}}_1$	7,206901	7,151619

Para este escenario bajo un diseño estratificado (ver cuadros 5,9 y 5,10) se obtienen por el estimador IV calibrado estimaciones poco sesgadas y muy eficientes, más aún resulta muy recomendable el uso del estimador en escenarios con pequeños tamaños muestrales ya que aun con variables instrumentales irrelevantes se obtienen buenas estimaciones.

		Sesgo relativo		Eficiencia relativa
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	1,001664	1,00161	0,9364169
	$\hat{\hat{B}}_1$	6,000276	6,000277	1,00407
Z_k irrelevante	$\hat{\hat{B}}_0$	0,4547473	0,5954126	0,550591
	$\hat{\hat{B}}_1$	6,206901	6,151619	0,5370082

Cuadro 5.9: Simulación $N = 10000$, $n = 200$

		Promedio de los coeficientes	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	2,007124	2,003918
	$\hat{\hat{B}}_1$	6,999929	7
Z_k irrelevante	$\hat{\hat{B}}_0$	1,075864	1,647381
	$\hat{\hat{B}}_1$	7,132983	7,053062

		Sesgo relativo		Eficiencia relativa
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	1,007124	1,003918	0,302455
	$\hat{\hat{B}}_1$	5,999929	6	$3{,}10384e-05$
Z_k irrelevante	$\hat{\hat{B}}_0$	0,07586359	0,6473815	0,1455922
	$\hat{\hat{B}}_1$	6,132983	6,053062	0,1592118

Cuadro 5.10: Simulación $N = 10000$, $n = 2000$

		Promedio de los coeficientes	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	2,003961	2,003721
	$\hat{\hat{B}}_1$	7,000127	7,000133
Z_k irrelevante	$\hat{\hat{B}}_0$	2,948267	3,024116
	$\hat{\hat{B}}_1$	6,863827	6,852

		Sesgo relativo		Eficiencia relativa
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	1,003961	1,003721	0,882552
	$\hat{\hat{B}}_1$	6,000127	6,000133	0,882552
Z_k irrelevante	$\hat{\hat{B}}_0$	1,948267	2,024116	1,16637
	$\hat{\hat{B}}_1$	5,863827	5,852	1,181248

Para el ultimo escenario se tiene un diseño con probabilidades proporcionales al tamaño πPT con tamaño muestral $n = 50$ no se observan grandes diferencias entre los dos estimadores. Sin embargo en el caso de tamaños muestrales más grandes, se logra notar que el estimador IV calibrado es más eficiente y presenta sesgos relativos pequeños o similares a los obtenidos por el estimador común.(ver cuadros 5,11, 5,12 y 5,13)

En general observamos que por medio del estimador propuesto se logran estimaciones mucho más eficientes que las del estimador común, siendo en algunos escenarios muy recomendable la aplicación de este.

Cuadro 5.11: Simulación $N = 10000$, $n = 50$

		Promedio de los coeficientes	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	2,123234	2,132091
	$\hat{\hat{B}}_1$	6,99733	6,997038
Z_k irrelevante	$\hat{\hat{B}}_0$	26,92713	17,54442
	$\hat{\hat{B}}_1$	3,255416	4,472071

		Sesgo relativo		Eficiencia relativa
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	1,123234	1,132091	1,148909
	$\hat{\hat{B}}_1$	5,99733	5,997038	1,148909
Z_k irrelevante	$\hat{\hat{B}}_0$	25,92713	16,54442	0,3888699
	$\hat{\hat{B}}_1$	2,255416	3,472071	0,4557456

Cuadro 5.12: Simulación $N = 10000$, $n = 100$

		Promedio de los coeficientes	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	2,534725	2,540073
	$\hat{\hat{B}}_1$	6,988553	6,988373
Z_k irrelevante	$\hat{\hat{B}}_0$	0,04877223	2,398323
	$\hat{\hat{B}}_1$	7,482183	7,05332

		Sesgo relativo		Eficiencia relativa
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	1,534725	1,540073	1,0201
	$\hat{\hat{B}}_1$	5,988553	5,988373	1,031812
Z_k irrelevante	$\hat{\hat{B}}_0$	−0,9512278	1,398323	0,04167293
	$\hat{\hat{B}}_1$	6,482183	6,05332	0,01222823

Cuadro 5.13: Simulación $N = 10000$, $n = 1000$

		Promedio de los coeficientes	
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	2,513922	2,513256
	$\hat{\hat{B}}_1$	6,989103	6,989123
Z_k irrelevante	$\hat{\hat{B}}_0$	1,323086	1,281082
	$\hat{\hat{B}}_1$	7,252157	7,244807

		Sesgo relativo		Eficiencia relativa
		$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
Z_k relevante	$\hat{\hat{B}}_0$	1,513922	1,513256	0,9974115
	$\hat{\hat{B}}_1$	5,989103	5,989123	0,9962215
Z_k irrelevante	$\hat{\hat{B}}_0$	0,3230861	0,2810819	1,127955
	$\hat{\hat{B}}_1$	6,252157	6,244807	0,9425528

Capítulo 6

Aplicación a situación real

Bonilla & Galvis (2012) estudian el impacto del grado de profesionalización de los docentes sobre los resultados en la calidad de la educación escolar medida por el desempeño académico de los estudiantes en la prueba saber 11, empleando modelos de variables instrumentales. En esta sección se utilizara el escenario planteado en este estudio, pero enfocándonos en el estimador por variables instrumentales y su eficiencia frente al estimador propuesto.

6.0.1. Descripción de los datos

La información que se analizara se obtuvo de los datos provistos por el DANE del formulario C-600 para el año 2009 y del ICFES de las pruebas saber 11 del año 2009 en los periodos 1 y 2.

Debido a que las instituciones educativas cuentan con varias sedes y diferentes jornadas, la unidad de análisis será la sede-jornada, de donde se calcula el número de estudiantes matriculados en secundaria y media, el número de profesores que enseñan en esos niveles, como el último nivel alcanzado por estos. Con esta información se construyen indicadores del grado de profesionalización de los docentes en cada sede, así como el indicador del número de estudiantes por profesor.

Las bases son cruzadas por nombre de sede, jornada y municipio, de estas se

descartan las jornadas nocturnas y fin de semana, finalmente se obtuvieron 3280 sede-jornadas con 72619 estudiantes.

Cuadro 6.1: Descripción datos

	Variable	N	Porcentaje
Grado de profesionalización	Profesionales o más	3280	59,08 %
	Posgrado	3280	14,60 %
	Formación pedagógica	3280	15,97 %
Institución	No oficial	3280	37,92 %
	Jornada completa	3280	33,47 %
	Genero mixto	3280	94,26 %

	Variable	N	Media o %
Individuo	Puntaje promedio	72619	41,58
	Puntaje lenguaje	72619	47,02
	Puntaje matemáticas	72619	45,31
	ln(Puntaje promedio)	72619	3,72
	ln(Puntaje lenguaje)	72619	0,172
	ln(Puntaje matemáticas)	72619	0,1782
	Edad	72619	16,8
	Personas hogar	72619	4,89
Educación Madre	Primaria	72619	33,48 %
	Secundaria	72619	41,64 %
	Técnico	72619	6,8 %
	Profesional	72619	11,41 %
	Posgrado	72619	2,5 %
Valor pensión	0	72619	58,36 %
	menos de 90000	72619	11,98 %
	Entre 90000 y 120000	72619	5,7 %
	Entre 120000 150000	72619	3,89 %
	Entre 150000 y 250000	72619	5,71 %
	250000 o más	72619	6,24 %

En los anteriores cuadros se muestra superficialmente como está conformada la población en estudio y las variables de interés que se usaran en la regresión.

6.0.2. Modelo e implementación

Al igual que en Bonilla & Galvis (2012) consideramos el siguiente modelo:

$$S_{11} = \beta_0 + \beta_1 \hat{T}_k + \beta_2 X_k + \varepsilon_k$$

donde S_{11} hace referencia a la variable (*promediodepuntaje*), T_k es la variable que indica el grado de profesionalización docente y X_k son variables de control como son los atributos del colegio y los atributos del individuo, dado que se supone la existencia de endogeneidad por medio de la variable T_k se procede conforme a la metodología de variables instrumentales teniendo por tanto como variable instrumental Z_k el número de estudiantes matriculados, sin embargo como lo explican los autores este instrumento puede fallar en el caso en que padres motivados por la calidad del estudio matriculen a sus hijos según sea más personalizada la enseñanza, para controlar este caso se usa la variable docentes por 100 alumnos.

En primera instancia se aplica el test de Hausman de exogeneidad, de donde tenemos que si el P-valor es $\prec$ 0,05 se rechaza la hipotesis de independencia o irrelevancia de las variables. Para esta prueba se obtuvo un p-valor de $7{,}309959e^{-37}$ por lo que se confirma la presencia de endogeneidad.

Ahora bien dadas las anteriores pautas se procede a la estimación del impacto bajo un diseño de muestreo MAS, donde el tamaño muestral es de $n = 1000$, se realiza un remuestreo sobre la población para observar el comportamiento del estimador, su sesgo y eficiencia frente al estimador común. A continuación se presentan los resultados obtenidos para los coeficiente β_0 y β_1.

Cuadro 6.2:

	Promedio de los coeficientes	
	$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$
$\hat{\hat{B}}_0$	47,11861	47,19863
$\hat{\hat{B}}_1$	0,3212795	0,3147185

Finalmente los resultados nos muestran que el estimador por variables instrumentales calibradas resulta ser más eficiente y presentar menos sesgo

	Sesgo relativo		Eficiencia relativa
	$\hat{\hat{B}}_{\pi,MC2E}$	$\hat{\hat{B}}_{calf,MC2E}$	
$\hat{\hat{B}}_0$	46,11861	46,19863	0,1524318
$\hat{\hat{B}}_1$	−0,6787205	−0,6852815	0,8009967

relativo, mostrándose como un estimador por así decirlo confiable, ya que el tamaño muestral implementado en realidad no es grande con respecto a la población en estudio y logra dar estimaciones más acertadas que el estimador común, más aun se podría considerar como recomendable su implementación en casos como este o como los mencionados anteriormente en la sección de simulación.

Capítulo 7

Conclusiones

El principal objetivo de esta investigación consistió en proponer un estimador para evaluación de impacto por variables instrumentales haciendo uso de la teoría de calibración, generando un conjunto de factores de expansión que modifica los factores originales inducidos por el diseño muestral.
Se realizaron simulaciones de Monte Carlo para distintos escenarios, en los cuales se comparó la eficiencia del estimador propuesto frente al estimador obtenido por la teoría clásica. Así mismo encontramos que el estimador calibrado presenta sesgos iguales o menores a los obtenidos por el estimador clásico, resultando más eficiente que este.

Bajo esta metodología en escenarios con T_k dicotómica en diseños muestrales MAS, se encontró que el estimador calibrado en pequeñas muestras obtiene estimaciones poco sesgadas pero no siempre más eficientes. Sin embargo en grandes muestras el estimador calibrado con primera etapa por regresión logística obtiene estimaciones menos sesgadas y mucho más eficientes que las estimaciones con el estimador común, aun en casos en los que la variable instrumental no cumple con la condición de relevancia, resultando en tales casos recomendable el uso del estimador propuesto.

En cuanto a los escenarios con T_k continua bajo los diseños MAS, estratificado MAS y πPT que el estimador IV calibrado obtiene estimaciones más eficientes que las obtenidas por el estimador clásico para tamaños de muestra grandes, aun cuando la variable instrumental no cumple con la

condición de relevancia. cabe resaltar que en muestras pequeñas el diseño estratificado MAS obtiene estimaciones con sesgo pequeño y mucho más eficientes que las del estimador común. Dados estos resultados resulta recomendable el uso de esta metodología para la estimación de impacto por IV.

Finalmente se observó en un escenario real el estimador por variables instrumentales calibrado, bajo el diseño MAS donde se obtuvieron estimadores, más eficientes y menos sesgados, en una muestra que se puede considerar pequeña dada la población en estudio y para tamaños de muestra más grandes se obtuvieron estimadores muy similares a los obtenidos con el estimador común. Bajo los planteamientos realizados en este trabajo queda abierto estudiar la metodología por el enfoque de calibración de minimización de distancia y su comparación con lo planteado en este trabajo.

Apéndice A

Esperanza y varianza del coeficiente $\hat{\hat{B}}$

En primera instancia consideramos la expresión

$$\hat{\hat{B}}_{calf} = \hat{\boldsymbol{T}}^{-1}_{calf,MC2E}\hat{\boldsymbol{t}}_{calf,MC2E}$$

donde

$$\hat{\boldsymbol{T}}_{calf,MC2E} = \boldsymbol{T}'\boldsymbol{P_z}\boldsymbol{w_k}\boldsymbol{T} \tag{A.1}$$

$$\hat{\boldsymbol{t}}_{calf,MC2E} = \boldsymbol{T}'\boldsymbol{P_z}\boldsymbol{w_k}\boldsymbol{y} \tag{A.2}$$

los elementos que componen estas matrices estarán respectivamente dados por

$$\hat{t}_{jj',calfMC2E} = \sum_s T_{jk}P_{z,jk}w_kT_{j'k}$$

$$\hat{t}_{j0,calfMC2E} = \sum_s T_{j'k}P_{z,jk}w_ky_k$$

Ahora bien hechas estas aclaraciones procedemos a tomar esperanza

$$\begin{aligned} E(\hat{\hat{B}}_{calf}) &= E\left[\sum_s T_{jk}P_{z,jk}w_kT_{j'k}\right] \\ &= \left(E\left(\hat{\boldsymbol{T}}_{calf,MC2E}\right)\right)^{-1} E\left(\hat{\boldsymbol{t}}_{calf,MC2E}\right) \end{aligned} \tag{A.3}$$

dado que tenemos matrices se tomará esperanza en cada elemento que las compone, por lo tanto tendremos que la esperanza $E\left(\hat{\boldsymbol{T}}_{calf,MC2E}\right)$ estará dada por

$$
\begin{aligned}
E(\hat{t}_{jj',calfMC2E}) &= E\left(\sum_s T_{jk}P_{z,jk}w_kT_{j'k}\right) \\
&= E\left(\sum_s T_{jk}P_{z,jk}(d_k+d_k\lambda z_k)T_{j'k}\right) \\
&= E\left(\sum_s T_{jk}P_{z,jk}d_kT_{j'k}+\sum_s T_{jk}P_{z,jk}d_k\lambda z_kT_{j'k}\right) \\
&= E\left(\hat{t}_{jj',\pi MC2E}\right)+E\left(\lambda\sum_s T_{jk}P_{z,jk}d_kz_kT_{j'k}\right) \\
&= t_{jj',MC2E}+E\left(\lambda\right)E\left(\sum_s T_{jk}P_{z,jk}d_kz_kT_{j'k}\right) \\
&= t_{jj',MC2E}+E\left((t_T-\hat{t}_{T,\pi})\left(\sum_s d_kz_kT_{j'k}\right)^{-1}\right) \\
&\quad E\left(\sum_s T_{jk}P_{z,jk}d_kz_kT_{j'k}\right) \\
&= t_{jj',MC2E}+(t_T-E(\hat{t}_{T,\pi}))\left(E\left(\sum_s d_kz_kT_{j'k}\right)\right)^{-1} \\
&\quad E\left(\sum_s T_{jk}P_{z,jk}d_kz_kT_{j'k}\right) \\
&= t_{jj',MC2E}+(t_T-t_T)\left(E\left(\sum_s d_kz_kT_{j'k}\right)\right)^{-1} \\
&\quad E\left(\sum_s T_{jk}P_{z,jk}d_kz_kT_{j'k}\right) \\
&= t_{jj',MC2E}+\cancel{(t_T-t_T)}\left(E\left(\sum_s d_kz_kT_{j'k}\right)\right)^{-1} \\
&\quad E\left(\sum_s T_{jk}P_{z,jk}d_kz_kT_{j'k}\right) \\
&= t_{jj',MC2E}+0 \qquad \text{(A.4)}
\end{aligned}
$$

Del resultado anterior podemos notar que $\hat{\boldsymbol{T}}_{calf,MC2E}$ es un estimador insesgado para $\boldsymbol{T}_{MC2E}$. Análogamente procedemos con $E\left(\hat{\boldsymbol{t}}_{calf,MC2E}\right)$

$$
\begin{aligned}
E(\hat{t}_{j0,calfMC2E}) &= E\left(\sum_s T_{j'k}P_{z,jk}w_k y_k\right) \\
&= E\left(\sum_s T_{j'k}P_{z,jk}(d_k + d_k\lambda z_k)y_k\right) \\
&= E\left(\sum_s T_{j'k}P_{z,jk}d_k y_k + \sum_s T_{j'k}P_{z,jk}d_k\lambda z_k y_k\right) \\
&= E\left(\hat{t}_{j0,\pi MC2E}\right) + E\left(\lambda\sum_s T_{j'k}P_{z,jk}d_k z_k y_k\right) \\
&= t_{j0,MC2E} + E\left(\lambda\right)E\left(\sum_s T_{j'k}P_{z,jk}d_k z_k y_k\right) \\
&= t_{j0,MC2E} + E\left((t_T - \hat{t}_{T,\pi})\left(\sum_s d_k z_k T_{j'k}\right)^{-1}\right) \\
&\quad E\left(\sum_s T_{j'k}P_{z,jk}d_k z_k y_k\right) \\
&= t_{j0,MC2E} + (t_T - E(\hat{t}_{T,\pi}))\left(E\left(\sum_s d_k z_k T_{j'k}\right)\right)^{-1} \\
&\quad E\left(\sum_s T_{j'k}P_{z,jk}d_k z_k y_k\right) \\
&= t_{j0,MC2E} + (t_T - t_T)\left(E\left(\sum_s d_k z_k T_{j'k}\right)\right)^{-1} \\
&\quad E\left(\sum_s T_{j'k}P_{z,jk}d_k z_k y_k\right) \\
&= t_{j0,MC2E} + \cancel{(t_T - t_T)}\left(E\left(\sum_s d_k z_k T_{j'k}\right)\right)^{-1} \\
&\quad E\left(\sum_s T_{j'k}P_{z,jk}d_k z_k y_k\right) \\
&= t_{j0,MC2E} + 0 \qquad \text{(A.5)}
\end{aligned}
$$

obteniendo igualmente que el estimador $\hat{\boldsymbol{t}}_{calf,MC2E}$ es insesgado para $\boldsymbol{t}_{MC2E}$, pero estos resultados no indican insesgades para el estimador $\hat{\hat{B}}_{calf}$.

Sin embargo (4,7) puede considerarse como una razón de totales, por consiguiente al aplicar los resultados expuestos en Särndal, Swensson & Wretman(2003) y Gutiérrez(2009) se tiene por linealización de Taylor la aproximación lineal

$$\hat{\hat{B}}_{calf} \doteq \hat{\hat{B}}_a = B + \sum_{j=1}^{q} \sum_{j' \leq j} a_{jj'} \left(\hat{t}_{jj',calfMC2E} - t_{jj',MC2E} \right) + \sum_{j=1}^{q} a_{j0} \left(\hat{t}_{j0,calfMC2E} - t_{j0,MC2E} \right) \qquad \text{(A.6)}$$

donde

$$a_{jj'} = \left. \frac{\partial \hat{\hat{B}}_{calf}}{\partial \hat{t}_{jj',calfMC2E}} \right|_{\hat{\boldsymbol{T}}_{calf,MC2E}=\boldsymbol{T}_{MC2E}, \hat{\boldsymbol{t}}_{calf,MC2E}=\boldsymbol{t}_{MC2E}}$$

$$a_{j0} = \left. \frac{\partial \hat{\hat{B}}_{calf}}{\partial \hat{t}_{j0,calfMC2E}} \right|_{\hat{\boldsymbol{T}}_{calf,MC2E}=\boldsymbol{T}_{MC2E}, \hat{\boldsymbol{t}}_{calf,MC2E}=\boldsymbol{t}_{MC2E}}$$

para los cuales las respectivas derivadas parciales estarán dadas por

$$\begin{aligned}
\frac{\partial \hat{\hat{B}}_{calf}}{\partial \hat{t}_{jj',calfMC2E}} &= \frac{\left(\partial \hat{\boldsymbol{T}}^{-1}_{calf,MC2E} \hat{\boldsymbol{t}}_{calf,MC2E} \right)}{\partial \hat{t}_{jj',calfMC2E}} \\
&= \frac{\left(\partial \hat{\boldsymbol{T}}^{-1}_{calf,MC2E} \right) \hat{\boldsymbol{t}}_{calf,MC2E}}{\partial \hat{t}_{jj',calfMC2E}} \\
&= \left(-\hat{\boldsymbol{T}}^{-1}_{calf,MC2E} \Lambda_{jj'} \hat{\boldsymbol{T}}^{-1}_{calf,MC2E} \right) \hat{\boldsymbol{t}}_{calf,MC2E} \\
&= -\hat{\boldsymbol{T}}^{-1}_{calf,MC2E} \Lambda_{jj'} \hat{\hat{B}}_{calf}
\end{aligned}$$

$$\begin{aligned}\frac{\partial \hat{\hat{B}}_{calf}}{\partial \hat{t}_{j0,calfMC2E}} &= \frac{\left(\partial \hat{\boldsymbol{T}}^{-1}_{calf,MC2E}\hat{\boldsymbol{t}}_{calf,MC2E}\right)}{\partial \hat{t}_{j0,calfMC2E}} \\ &= \hat{\boldsymbol{T}}^{-1}_{calf,MC2E}\frac{\left(\partial \hat{\boldsymbol{t}}_{calf,MC2E}\right)}{\partial \hat{t}_{j0,calfMC2E}} \\ &= \hat{\boldsymbol{T}}^{-1}_{calf,MC2E}\lambda_j\end{aligned}$$

donde $\Lambda_{jj'}$ es una matriz que en las posiciones (j, j') y (j', j) tiene unos y en las otras tiene ceros, mientras que λ_j es un vector que en los componentes j tiene unos y ceros en otro caso.

Ahora bien para obtener la forma linealizada de $\hat{\hat{B}}_{calf}$ se procede a evaluar las derivadas en el punto $(\boldsymbol{T}_{MC2E}, \boldsymbol{t}_{MC2E})$ y se insertan en (A,6) de modo que

$$\begin{aligned}\hat{\hat{B}}_a &= \boldsymbol{B} - \sum_{j=1}^{q}\sum_{j'\leq j}\boldsymbol{T}^{-1}_{MC2E}\Lambda_{jj'}\boldsymbol{B}\left(\hat{t}_{jj',calfMC2E} - t_{jj',MC2E}\right) + \\ &\quad \sum_{j=1}^{q}\boldsymbol{T}^{-1}_{MC2E}\lambda_j\left(\hat{t}_{j0,calfMC2E} - t_{j0,MC2E}\right) \\ &= \boldsymbol{B} - \boldsymbol{T}^{-1}_{MC2E}\left(\hat{\boldsymbol{T}}_{calf,MC2E} - \boldsymbol{T}_{MC2E}\right)\boldsymbol{B} + \\ &\quad \boldsymbol{T}^{-1}_{MC2E}\left(\hat{\boldsymbol{t}}_{calf,MC2E} - \boldsymbol{t}_{MC2E}\right) \\ &= \boldsymbol{B} + \boldsymbol{T}^{-1}_{MC2E}\left(\hat{\boldsymbol{t}}_{calf,MC2E} - \hat{\boldsymbol{T}}_{calf,MC2E}\boldsymbol{B}\right) \qquad \text{(A.7)}\end{aligned}$$

tomamos esperanza de la anterior expresión

$$\begin{aligned}E(\hat{\hat{B}}_a) &= E\left(\boldsymbol{B} + \boldsymbol{T}^{-1}_{MC2E}\left(\hat{\boldsymbol{t}}_{calf,MC2E} - \hat{\boldsymbol{T}}_{calf,MC2E}\boldsymbol{B}\right)\right) \\ &= \boldsymbol{B} + \boldsymbol{T}^{-1}_{MC2E}\left(E\left(\hat{\boldsymbol{t}}_{calf,MC2E}\right) - E\left(\hat{\boldsymbol{T}}_{calf,MC2E}\right)\boldsymbol{B}\right) \\ &= \boldsymbol{B} + \boldsymbol{T}^{-1}_{MC2E}\left(\boldsymbol{t}_{MC2E} - \boldsymbol{T}_{MC2E}\boldsymbol{B}\right) \\ &= \boldsymbol{B} + \boldsymbol{T}^{-1}_{MC2E}\cancel{\left(\boldsymbol{t}_{MC2E} - \boldsymbol{t}_{MC2E}\right)} \qquad \text{(A.8)} \\ &= \boldsymbol{B} + 0\end{aligned}$$

finalmente por propiedades podemos decir que el estimador $\hat{\hat{B}}_{calf}$ es aproximadamente insesgado para B.

Considerando las afirmaciones anteriores y como lo ilustran Särndal, Swensson & Wretman(2003) se puede llegar a una aproximación de la matriz de varianzas y covarianzas

$$\begin{aligned}
AV(\hat{\hat{B}}_{calf}) &= V\left(\boldsymbol{B} + \boldsymbol{T}_{MC2E}^{-1}\left(\hat{\boldsymbol{t}}_{calf,MC2E} - \hat{\boldsymbol{T}}_{calf,MC2E}\boldsymbol{B}\right)\right) \\
&= \cancel{V(\boldsymbol{B})} + V\left(\boldsymbol{T}_{MC2E}^{-1}\left(\hat{\boldsymbol{t}}_{calf,MC2E} - \hat{\boldsymbol{T}}_{calf,MC2E}\boldsymbol{B}\right)\right) \\
&= 0 + \boldsymbol{T}_{MC2E}^{-1} V\left(\hat{\boldsymbol{t}}_{calf,MC2E} - \hat{\boldsymbol{T}}_{calf,MC2E}\boldsymbol{B}\right)\boldsymbol{T}_{MC2E}^{-1} \\
&= \boldsymbol{T}_{MC2E}^{-1} V\left(\boldsymbol{T}'\boldsymbol{P}_{\boldsymbol{z}}\boldsymbol{w}_{\boldsymbol{k}}\boldsymbol{Y} - \boldsymbol{T}'\boldsymbol{P}_{\boldsymbol{z}}\boldsymbol{w}_{\boldsymbol{k}}\boldsymbol{T}\boldsymbol{B}\right)\boldsymbol{T}_{MC2E}^{-1} \\
&= \boldsymbol{T}_{MC2E}^{-1} V\left(\boldsymbol{T}'\boldsymbol{P}_{\boldsymbol{z}}\boldsymbol{w}_{\boldsymbol{k}}\left(\boldsymbol{Y} - \boldsymbol{T}\boldsymbol{B}\right)\right)\boldsymbol{T}_{MC2E}^{-1} \\
&= \boldsymbol{T}_{MC2E}^{-1} V\left(\boldsymbol{T}'\boldsymbol{P}_{\boldsymbol{z}}\boldsymbol{w}_{\boldsymbol{k}}\boldsymbol{E}_{\boldsymbol{k}}\right)\boldsymbol{T}_{MC2E}^{-1} \\
&= \boldsymbol{T}_{MC2E}^{-1}\boldsymbol{V}\boldsymbol{T}_{MC2E}^{-1}
\end{aligned}$$

de lo anterior tenemos que, $\boldsymbol{V}$ expresa una matriz de varianzas y covarianzas de un estimador de calibración por enfoque funcional donde el parámetro de interés está dado por $\boldsymbol{T}'\boldsymbol{P}_{\boldsymbol{z}}\boldsymbol{E}_{\boldsymbol{k}}$ y $\boldsymbol{E}_{\boldsymbol{k}}$ hace referencia al vector que contiene los residuales de la regresión ajustada. Bajo este planteamiento, consideramos las expresiones dadas en Särndal & Estevao(2000) para la varianza de un estimador de calibración, por ende se tiene que los elementos de la matriz estarán dados por

$$v_{jj'} = \sum\sum_{U} D_{kl} H_k H_l \tag{A.9}$$

De donde se tiene que

$$\begin{array}{llll}
d_{kl} = \frac{1}{\pi_{kl}} & D_{kl} = \frac{d_k d_l}{d_{kl}-1} & \text{para } k \neq l \\
d_{kl} = d_k & D_{kl} = d_k - 1 & \text{para } k = l
\end{array}$$

$$\begin{aligned}
H_k &= T_{jk} P_{zjk} E_k - T_{j'k} Q \\
H_l &= T_{jl} P_{zjl} E_l - T_{j'l} Q
\end{aligned} \tag{A.10}$$

con

$$Q = \left(\sum_U z_k T_{j'k} \right)^{-1} \sum_U z_k T_{jk} P_{zjk} E_k \tag{A.11}$$

El estimador de la matriz de varianzas y covarianzas es

$$\hat{V}(\hat{\hat{B}}_{calf}) = \left(\sum_s T_{jk} P_{z,jk} w_k T_{j'k} \right)^{-1} \hat{\boldsymbol{V}} \left(\sum_s T_{jk} P_{z,jk} w_k T_{j'k} \right)^{-1} \tag{A.12}$$

Donde los elementos de la matriz simétrica $\hat{\boldsymbol{V}}$ son

$$\hat{v_{jj'}} = \sum \sum_s D_{kl} h_k h_l \tag{A.13}$$

De donde se tiene que

$$h_k = T_{jk} P_{zjk} e_k - T_{j'k} \hat{Q} \tag{A.14}$$

con

$$\hat{Q} = \left(\sum_s d_k z_k T_{j'k} \right)^{-1} \sum_s d_k z_k T_{jk} P_{zjk} e_k \tag{A.15}$$

y e_k es el residual ajustado muestral,

$$\hat{e}_k = y_k - T_{j'k} \hat{\hat{B}}_{calf}$$

Apéndice B

Código R simulación

B.1. Diseño MAS con T_k dicotómica

```
install.packages("dplyr")
install.packages("stats")
install.packages("tidyr")
install.packages("TeachingSampling")
install.packages("ggplot2")
install.packages("ivpack")
#install.packages("sqldf")
#install.packages("survey")

library(dplyr)
library(stats)
library(tidyr)
library(TeachingSampling)
library(ggplot2)
library(ivpack)
library(graphics)

####   Modelo Teorico ##############################
Epsilon=rnorm(mean=0,sd=2)
Gamma0= 1
Gamma1= 0.9
```

```
ZK<-rnorm(mean= 3, sd= 5)
WK=rnorm(mean=0,sd=1)
Link=exp(1+0.9*ZK+WK+Epsilon)/(1+exp(1+0.9*ZK+WK+Epsilon))
# TK variable dicotomica que se obtiene a partir de la funcion link
Beta0=3
Beta1=9
YK= 3 + 9*TK + Epsilon
#####################################################################
#####################################################################
######################### Modelo Poblacional ########################
set.seed(201606)
N <- 10000 ###tamaño poblalcion finita

Ek<-rnorm(N)##error o residual del modelo
Zk<-rnorm(N,mean= 3, sd= 5)##variable instrumental
G0=1
G1=0.01
p <- exp(G0+G1*Zk+8*Ek)/(1+exp(G0+G1*Zk+8*Ek)) ##funcion link
Tk<-rbinom(N,1,p) ###variable indicadora de tratamiento
poblacion<-data.frame(Zk)
poblacion$Tk<-Tk
B0=2
B1=7
poblacion$Yk<-B0+B1*poblacion$Tk+Ek ####variable de interes

#####Modelos poblacionales   ##############################
Modelo1=lm(Yk~Tk, data=poblacion)
###se revisa en el modelo la relevancia  de Zk
ModeloTk=glm(Tk~Zk,family=quasibinomial,  data=poblacion)
summary(ModeloTk)
cor(Zk,Ek)###exogeneidad del instrumento
Tk.h<-predict(ModeloTk, type="response")
poblacion$Tk.h<-Tk.h
MCO<-lm(poblacion$Yk~Tk.h)
###Modelo con funcion ivreg
IV<-ivreg(Yk~Tk|Zk, data = poblacion)
```

```
#### Identificacion de la poblacion en tratamientos y controles
PT<-filter(poblacion, Tk==1)
PC<-filter(poblacion, Tk==0)

#####################################################################
#####################################################################
#################### Estimacion en una muestra ###############
calmues<-function(){
n=1000###tamano muestral

######Seleccion de muestra
sam<- S.SI(N, n)
datasam<- poblacion[sam, ]
datasam$ak <- rep(N / n, rep = n)
############## Modelos con pesos diseño MAS
####modelo logit
modTk<-glm(Tk~Zk,family=quasibinomial, weights= ak, data=datasam)
Tk.hat <- predict(modTk, type="response")
datasam$Tk.hat<-Tk.hat
mody<-lm(Yk~Tk.hat, weights= ak, data=datasam)
mdIV<-ivreg(Yk~Tk|Zk,weights= ak, data=datasam)

##########################calibracion
######Supuestos
Nt=length(PT$Yk)
Nc=length(PC$Yk)
###poblacion tratamiento y control en la muestra
Ptm<-filter(datasam, Tk==1)
Pcm<-filter(datasam, Tk==0)
nt=length(Ptm$Yk)
nc=length(Pcm$Yk)

##Estimacion del total para la variable TK
##por HT en tratamientos y controles
Ttht=sum(Ptm$ak)
Tcht=sum(Pcm$ak)
```

```
#Estimacion de pesos calf para tratamientos
TZkt=Ptm$Zk*t(Ptm$Tk.hat)
yZkt= Ptm$Zk*Ptm$Yk
wkt=(Nt/nt)*(1+(abs(Nt-Ttht)*((Nt/nt)*sum(TZkt))^(-1))*Ptm$Zk)
Ptm$wk<-wkt
###Controles
TZkc=Pcm$Zk*t(1-Pcm$Tk.hat)
yZkc= Pcm$Zk*Pcm$Yk
wkc=(Nc/nc)*(1+(abs(Tcht-Nc)*((Nc/nc)*sum(TZkc))^(-1))*Pcm$Zk)
Pcm$wk<-wkc
data<-union(Ptm,Pcm)

##Forma matricial
xb0<-rep(1,n)
data$xb0<-xb0
xx=matrix(c(data$xb0,data$Tk),ncol=2,nrow=n)
xx2=matrix(c(data$xb0,data$Tk.hat),ncol=2,nrow=n)
Zz=matrix(c(data$xb0,data$Zk),ncol=2, nrow=n)
Pz=Zz%*%solve(t(Zz)%*%Zz)%*%t(Zz)
##regresion con MC2E
akd<-diag(x=data$ak,ncol=n,nrow=n) ###matriz pesos diseño MAS
wkd<-diag(x=data$wk,ncol=n,nrow=n) ###matriz pesos calibrados
##estimacion H-T
betht=(solve(t(xx)%*%Pz%*%akd%*%xx))%*%t(xx)%*%Pz%*%akd%*%data$Yk
##estimacion calibrada
betcalf=(solve(t(xx)%*%Pz%*%wkd%*%xx))%*%t(xx)%*%Pz%*%wkd%*%data$Yk
###regresion con  pseudo verosimilitud y Mco
###estimacion H-T
betht2=(solve(t(xx2)%*%akd%*%xx2))%*%t(xx2)%*%akd%*%data$Yk
###estimacion calibrada
betcalf2=(solve(t(xx2)%*%wkd%*%xx2))%*%t(xx2)%*%wkd%*%data$Yk
##betas calibrados y H-T
beta0h<-data.frame(betht[1])
beta1h<-data.frame(betht[2])
beta0c<-data.frame(betcalf[1])
beta1c<-data.frame(betcalf[2])
```

```
beta0h1<-data.frame(betht2[1])
beta1h1<-data.frame(betht2[2])
beta0c1<-data.frame(betcalf2[1])
beta1c1<-data.frame(betcalf2[2])
####varianza
ekcalf=data$Yk-xx%*%betcalf
par<-as.numeric(c(beta0h,beta0c,beta1h,beta1c,
beta0h1,beta0c1,beta1h1,beta1c1))
par
}
nsim=1000
simST <- replicate(nsim, calmues())
betST0h<-simST[1,1:nsim]
betST0c<-simST[2,1:nsim]
betST1h<-simST[3,1:nsim]
betST1c<-simST[4,1:nsim]
betST0h1<-simST[5,1:nsim]
betST0c1<-simST[6,1:nsim]
betST1h1<-simST[7,1:nsim]
betST1c1<-simST[8,1:nsim]
###medias de los estimadores segun las n simulaciones
mean(betST0h)
mean(betST0c)
mean(betST1h)
mean(betST1c)
mean(betST0h1)
mean(betST0c1)
mean(betST1h1)
mean(betST1c1)
####Sesgo relativo de los betas
####betas con MC2E
Br0h=(1/nsim)*sum(betST0h-B0/B0)
Br1h=(1/nsim)*sum(betST1h-B1/B1)
##betas con MC2E
Br0h1=(1/nsim)*sum(betST0h1-B0/B0)
Br1h1=(1/nsim)*sum(betST1h1-B1/B1)
```

```
##betas con MC2E
Br0=(1/nsim)*sum(betST0c-B0/B0)
Br1=(1/nsim)*sum(betST1c-B1/B1)
#betas con pseudoverosimilitud y mco
Br01=(1/nsim)*sum(betST0c1-B0/B0)
Br11=(1/nsim)*sum(betST1c1-B1/B1)
##Eficiencia relativa
ER0=sum(betST0c-B0)^2/sum(betST0h-B0)^2
ER1=sum(betST1c-B1)^2/sum(betST1h-B1)^2
ER01=sum(betST0c1-B0)^2/sum(betST0h1-B0)^2
ER11=sum(betST1c1-B1)^2/sum(betST1h1-B1)^2
```

B.2. Diseño estratificado MAS con T_k dicotómica

```
set.seed(201606)
N <- 10000

Ek<-rnorm(N)
Zk<-rnorm(N,mean= 3, sd= 5)
G0=1
G1=0.01
p <- exp(G0+G1*Zk+9*Ek)/(1+exp(G0+G1*Zk+9*Ek))
Tk<-rbinom(N,1,p)
poblacion<-data.frame(Zk)
poblacion$Tk<-Tk
B0=2
B1=7
poblacion$Yk<-B0+B1*poblacion$Tk+Ek

#####Modelos poblacionales
Modelo1=lm(Yk~Tk, data=poblacion)

ModeloTk=glm(Tk~Zk,family=quasibinomial,  data=poblacion)
summary(ModeloTk)
Tk.h<-predict(ModeloTk, type="response")
poblacion$Tk.h<-Tk.h
```

```
MCO2<-lm(poblacion$Yk~Tk.h)

###Modelo con funcion ivreg
IV<-ivreg(Yk~Tk|Zk, data = poblacion)

PT<-filter(poblacion, Tk==1)
PC<-filter(poblacion, Tk==0)
#Generacion de Estratos
Estrato<-character(length(PT$Yk))
for(i in 1:length(PT$Yk)) if
(PT$Yk[i]<=8.8) Estrato[i]<-"Estrato1"
else Estrato[i]<- "Estrato2"
PT$Estrato<-as.factor(Estrato)
Estrato<-character(length(PC$Yk))
for(i in 1:length(PC$Yk)) if
(PC$Yk[i]<2) Estrato[i]<-"Estrato1"
else Estrato[i]<- "Estrato2"
PC$Estrato<-as.factor(Estrato)
poblacion<-union(PT,PC)

###Estimacion en una muestra ####
###tamaño muestral
Ttpob=length(PT$Tk)
Tcpob=length(PC$Tk)
Nt=Ttpob
Nt1<-summary(PT$Estrato)[[1]]
Nt2<-summary(PT$Estrato)[[2]]
Nth=c(Nt1,Nt2)
Nc=Tcpob
Nc1<-summary(PC$Estrato)[[1]]
Nc2<-summary(PC$Estrato)[[2]]
Nch=c(Nc1,Nc2)
###tamanos muestrales
calmuES<-function(){
nt1=100
nt2=900
```

```
nth=c(nt1,nt2)
nc1=890
nc2=110
nch=c(nc1,nc2)

##Seleccion de muestras en los grupos tratamiento y control
samt <- S.STSI(PT$Estrato, Nth, nth)
Ptmues<- PT[samt, ]
#pesos ak en tratamientos
ak<-numeric(length(Ptmues$Estrato))
for (i in 1:length(Ptmues$Estrato)) if
(Ptmues$Estrato[i]== "Estrato1") ak[i] <-Nt1/nt1
else ak[i]<- Nt2/nt2
Ptmues$ak<-ak
samc <- S.STSI(PC$Estrato, Nch, nch)
Pcmues<- PC[samc, ]
##pesos ak en controles
ak<-numeric(length(Pcmues$Estrato))
for (i in 1:length(Pcmues$Estrato)) if
(Pcmues$Estrato[i]== "Estrato1") ak[i] <-Nc1/nc1
else ak[i]<- Nc2/nc2
Pcmues$ak<-ak
datasam<-union(Ptmues,Pcmues)

#Generacion de variables indicadoras de los estratos
#en la muestra
Es1<-numeric(length(datasam$Estrato))
for (i in 1:length(datasam$Estrato)) if
(datasam$Estrato[i]== "Estrato1") Es1[i] <- 1 else Es1[i]<- 0
datasam$Es1<-Es1

Es2<-numeric(length(datasam$Estrato))
for (i in 1:length(datasam$Estrato)) if
(datasam$Estrato[i]== "Estrato2") Es2[i] <- 1
else Es2[i]<- 0
datasam$Es2<-Es2
```

```
### Modelos con pesos diseño MAS

mTk<-glm(Tk~Zk,family=quasibinomial, weights= ak, data=datasam)
Tk.hat <- predict(mTk, type="response")
datasam$Tk.hat<-Tk.hat
mody<-lm(Yk~Tk.hat, weights= ak, data=datasam)
mdIV<-ivreg(Yk~Tk|Zk,weights= ak, data=datasam)

###calibracion
#Supuestos
Nt1
Nt2
Nc1
Nc2
#poblacion tratamiento y control en la muestra
Ptm<-filter(datasam, Tk==1)
Pcm<-filter(datasam, Tk==0)

##Estimacion del total para la variable TK.hat
## por HT en tratamientos
##y controles en cada estrato
Ttht1=sum(Ptm$ak*Ptm$Es1*Ptm$Tk.hat)
Ttht2=sum(Ptm$ak*Ptm$Es2*Ptm$Tk.hat)
Tcht1=sum(Pcm$ak*Pcm$Es1*(1-Pcm$Tk.hat))
Tcht2=sum(Pcm$ak*Pcm$Es2*(1-Pcm$Tk.hat))
##Estimación de pesos calf para tratamientos
Ptm1<-filter(Ptm, Es1==1)
Ptm2<-filter(Ptm, Es2==1)
TZkt1=Ptm1$Zk*t(Ptm1$Tk.hat)
yZkt1= Ptm1$Zk*Ptm1$Yk
wkt1=(Nt1/nt1)*(1+((Nt1-Ttht1)*((Nt1/nt1)*sum(TZkt1))^(-1))*Ptm1$Zk)
Ptm1$wk<-wkt1

TZkt2=Ptm2$Zk*t(Ptm2$Tk.hat)
yZkt2= Ptm2$Zk*Ptm2$Yk
wkt2=(Nt2/nt2)*(1+((Nt2-Ttht2)*((Nt2/nt2)*sum(TZkt2))^(-1))*Ptm2$Zk)
```

```
Ptm2$wk<-wkt2
Ptm<-union(Ptm1,Ptm2)
####Controles
Pcm1<-filter(Pcm, Es1==1)
Pcm2<-filter(Pcm, Es2==1)
TZkc1=Pcm1$Zk*t(1-Pcm1$Tk.hat)
yZkc1= Pcm1$Zk*Pcm1$Yk
wkc1=(Nc1/nc1)*(1+((Tcht1-Nc1)*((Nc1/nc1)*sum(TZkc1))^(-1))*Pcm1$Zk)
Pcm1$wk<-wkc1
TZkc2=Pcm2$Zk*t(1-Pcm2$Tk.hat)
yZkc2= Pcm2$Zk*Pcm2$Yk
wkc2=(Nc2/nc2)*(1+((Tcht2-Nc2)*((Nc2/nc2)*sum(TZkc2))^(-1))*Pcm2$Zk)
Pcm2$wk<-wkc2
Pcm<-union(Pcm1,Pcm2)

data<-union(Ptm,Pcm)
n<-length(data$Yk)

##Forma matricial
xb0<-rep(1,n)
data$xb0<-xb0
xx=matrix(c(data$xb0,data$Tk),ncol=2,nrow=n)
xx2=matrix(c(data$xb0,data$Tk.hat),ncol=2,nrow=n)
Zz=matrix(c(data$xb0,data$Zk),ncol=2, nrow=n)
Pz=Zz%*%solve(t(Zz)%*%Zz)%*%t(Zz)

##regresion con MC2E
akd<-diag(x=data$ak,ncol=n,nrow=n)
wkd<-diag(x=data$wk,ncol=n,nrow=n)
bethtES=(solve(t(xx)%*%Pz%*%akd%*%xx))%*%t(xx)%*%Pz%*%akd%*%data$Yk
betcalfES=(solve(t(xx)%*%Pz%*%wkd%*%xx))%*%t(xx)%*%Pz%*%wkd%*%data$Yk
##regresion con  pseudo verosimilitud y Mco
bethtES2=(solve(t(xx2)%*%akd%*%xx2))%*%t(xx2)%*%akd%*%data$Yk
betcalfES2=(solve(t(xx2)%*%wkd%*%xx2))%*%t(xx2)%*%wkd%*%data$Yk

##betas calibrados y H-T
```

```
beta0hES<-data.frame(bethtES[1])
beta1hES<-data.frame(bethtES[2])
beta0cES<-data.frame(betcalfES[1])
beta1cES<-data.frame(betcalfES[2])
beta0h1ES2<-data.frame(bethtES2[1])
beta1h1ES2<-data.frame(bethtES2[2])
beta0c1ES2<-data.frame(betcalfES2[1])
beta1c1ES2<-data.frame(betcalfES2[2])
par<-as.numeric(c(beta0hES,beta0cES,beta1hES,beta1cES,beta0h1ES2
,beta0c1ES2,beta1h1ES2,beta1c1ES2))
par
}

nsim=1000
simST <- replicate(nsim, calmuES())

beta0hES<-simST[1,1:nsim]
beta0cES<-simST[2,1:nsim]
beta1hES<-simST[3,1:nsim]
beta1cES<-simST[4,1:nsim]
beta0h1ES2<-simST[5,1:nsim]
beta0c1ES2<-simST[6,1:nsim]
beta1h1ES2<-simST[7,1:nsim]
beta1c1ES2<-simST[8,1:nsim]
mean(beta0hES)
mean(beta0cES)
mean(beta1hES)
mean(beta1cES)
mean(beta0h1ES2)
mean(beta0c1ES2)
mean(beta1h1ES2)
mean(beta1c1ES2)
###Sesgo relativo de los betas
Br0h=(1/nsim)*sum(beta0hES-B0/B0)##betas con MC2E
Br1h=(1/nsim)*sum(beta1hES-B1/B1)
Br0h1=(1/nsim)*sum(beta0h1ES2-B0/B0)##betas con MC2E
```

```
Br1h1=(1/nsim)*sum(beta1h1ES2-B1/B1)

Br0=(1/nsim)*sum(beta0cES-B0/B0)##betas con MC2E
Br1=(1/nsim)*sum(beta1cES-B1/B1)
##betas con pseudoverosimilitud y mco
Br01=(1/nsim)*sum(beta0c1ES2-B0/B0)
Br11=(1/nsim)*sum(beta1c1ES2-B1/B1)
#####Eficiencia relativa
ER0=sum(beta0cES-B0)^2/sum(beta0hES-B0)^2
ER1=sum(beta1cES-B1)^2/sum(beta1hES-B1)^2
ER01=sum(beta0c1ES2-B0)^2/sum(beta0h1ES2-B0)^2
ER11=sum(beta1c1ES2-B1)^2/sum(beta1h1ES2-B1)^2
```

B.3. Diseño MAS con T_k continua

```
set.seed(201606)
N <- 10000
v=rnorm(N,mean=2,sd=2)
Ek<-rnorm(N)
Zk<-rnorm(N,mean= 3, sd= 5)
G0=0.7
G1=0.01
IT<-rbinom(N,1,0.5)
Tk<-G0+G1*Zk+3.8*Ek+v
poblacion<-data.frame(Zk)
poblacion$Tk<-Tk
poblacion$IT<-IT
B0=2
B1=7
poblacion$Yk<-B0+B1*poblacion$Tk+Ek
#####Modelos poblacionales
Modelo1=lm(Yk~Tk, data=poblacion)
mode=lm(Tk~Zk, data=poblacion)
summary(mode)####condicion de relevancia
cor(Ek,Zk)###Condicion de exogeneidad
```

```
###Modelo con funcion ivreg
IV<-ivreg(Yk~Tk|Zk, data = poblacion)
#### Identificacion de la poblacion en tratamientos y controles
IT2<-numeric(length(poblacion$IT))
for (i in 1:length(poblacion$IT)) if
(poblacion$IT[i]== 1) IT2[i] <- "T" else IT2[i]<- "C"
poblacion$IT<-IT2

PT<-filter(poblacion, IT=="T")
PC<-filter(poblacion, IT=="C")
cpobla=sum(PC$Tk)
Tpobla=sum(PT$Tk)
#########################################################
#########################################################
######## Estimacion en una muestra ##################

###tamaño muestral
calmuc<-function(){
n=1000
######Seleccion de muestra
sam<- S.SI(N, n)
datasam<- poblacion[sam, ]
datasam$ak <- rep(N / n, rep = n)

############ Modelos con pesos diseño MAS

modTk<-lm(Tk~Zk, weights= ak, data=datasam)
Tk.hat <- predict(modTk, type="response")
datasam$Tk.hat<-Tk.hat
mdIV<-ivreg(Yk~Tk|Zk,weights= ak, data=datasam)
############calibracion ########################
######Supuestos
Nt=length(PT$Yk)
Nc=length(PC$Yk)
##poblacion tratamiento y control en la muestra
Ptm<-filter(datasam, IT=="T")
```

```
Pcm<-filter(datasam, IT=="C")
nt=length(Ptm$Yk)
nc=length(Pcm$Yk)

####Estimacion del total para la variable TK.hat por
##HT en tratamientos y controles
Ttht=sum(Ptm$ak*Ptm$Tk.hat)
Tcht=sum(Pcm$ak*Pcm$Tk.hat)
#Estimacion de pesos calf para tratamientos
TZkt=Ptm$Zk*t(Ptm$Tk.hat)
yZkt= Ptm$Zk*Ptm$Yk
wkt=(Nt/nt)*(1+((Tpobla-Ttht)*((Nt/nt)*sum(TZkt))^(-1))*Ptm$Zk)
Ptm$wk<-wkt

###Controles
TZkc=Pcm$Zk*t(Pcm$Tk.hat)
yZkc= Pcm$Zk*Pcm$Yk
wkc=(Nc/nc)*(1+((Tcht-cpobla)*((Nc/nc)*sum(TZkc))^(-1))*Pcm$Zk)
Pcm$wk<-wkc
data<-union(Ptm,Pcm)
#########Estimacion de Bcalf
xb0<-rep(1,n)
data$xb0<-xb0
xx=matrix(c(data$xb0,data$Tk),ncol=2,nrow=n)
xx2=matrix(c(data$xb0,data$Tk.hat),ncol=2,nrow=n)
Zz=matrix(c(data$xb0,data$Zk),ncol=2, nrow=n)
Pz=Zz%*%solve(t(Zz)%*%Zz)%*%t(Zz)
#########regresion con MC2E
akd<-diag(x=data$ak,ncol=n,nrow=n)
bethtc=(solve(t(xx)%*%Pz%*%akd%*%xx))%*%t(xx)%*%Pz%*%akd%*%data$Yk

wkd<-diag(x=data$wk,ncol=n,nrow=n)
betcalfc=(solve(t(xx)%*%Pz%*%wkd%*%xx))%*%t(xx)%*%Pz%*%wkd%*%data$Yk

beta0hc<-data.frame(bethtc[1])
beta1hc<-data.frame(bethtc[2])
```

```
beta0cc<-data.frame(betcalfc[1])
beta1cc<-data.frame(betcalfc[2])

par<-as.numeric(c(beta0hc,beta0cc,beta1hc,beta1cc))
par
}
nsim=1000
simSTc <- replicate(nsim, calmuc())
betST0hc<-simSTc[1,1:nsim]
betST0cc<-simSTc[2,1:nsim]
betST1hc<-simSTc[3,1:nsim]
betST1cc<-simSTc[4,1:nsim]

mean(betST0hc)
mean(betST0cc)
mean(betST1hc)
mean(betST1cc)
###############Sesgo relativo de los betas
Br0hc=(1/nsim)*sum(betST0hc-B0/B0)####betas con HT
Br1hc=(1/nsim)*sum(betST1hc-B1/B1)
Br0c=(1/nsim)*sum(betST0cc-B0/B0)####betas con calf
Br1c=(1/nsim)*sum(betST1cc-B1/B1)
###############Eficiencia relativa
ER0c=sum(betST0cc-B0)^2/sum(betST0hc-B0)^2
ER1c=sum(betST1cc-B1)^2/sum(betST1hc-B1)^2
```

B.4. Diseño estratificado MAS con T_k continua

```
#################################################################
################## Modelo Poblacional ###########################
set.seed(201606)
N <- 10000

v=rnorm(N,mean=0,sd=4)
Ek<-rnorm(N)
Zk<-rnorm(N,mean= 5, sd= 7)
```

```
G0=7
G1=0.01
IT<-rbinom(N,1,0.5)
Tk<-G0+G1*Zk+8*Ek+v
poblacion<-data.frame(Zk)
poblacion$Tk<-Tk
poblacion$IT<-IT
B0=2
B1=7
poblacion$Yk<-B0+B1*poblacion$Tk+Ek
#####Modelos poblacionales
Modelo1=lm(Yk~Tk, data=poblacion)

ModeloTk=lm(Tk~Zk,  data=poblacion)
summary(ModeloTk)###relevancia del instrumento
cor(Ek,Zk)###condicion de exogeneidad
Tk.h<-predict(ModeloTk, type="response")
poblacion$Tk.h<-Tk.h
###Modelo con funcion ivreg
IV<-ivreg(Yk~Tk|Zk, data = poblacion)
#### Identificacion de la poblacion en tratamientos y controles
IT2<-numeric(length(poblacion$IT))
for (i in 1:length(poblacion$IT)) if
(poblacion$IT[i]== 1) IT2[i] <- "T" else IT2[i]<- "C"
poblacion$IT<-IT2

PT<-filter(poblacion, IT=="T")
PC<-filter(poblacion, IT=="C")
cpobla=sum(PC$Tk)
Tpobla=sum(PT$Tk)

#Generacion de Estratos
Estrato<-character(length(PT$Yk))
for(i in 1:length(PT$Yk)) if
(PT$Yk[i]<=8.8) Estrato[i]<-"Estrato1" else Estrato[i]<- "Estrato2"
PT$Estrato<-as.factor(Estrato)
```

```
Estrato<-character(length(PC$Yk))
for(i in 1:length(PC$Yk)) if
(PC$Yk[i]<2) Estrato[i]<-"Estrato1" else Estrato[i]<- "Estrato2"
PC$Estrato<-as.factor(Estrato)

poblacion<-union(PT,PC)

##Generacion de variables indicadoras de los estratos en la muestra
Es1<-numeric(length(poblacion$Estrato))
for (i in 1:length(poblacion$Estrato)) if
(poblacion$Estrato[i]== "Estrato1") Es1[i] <- 1 else Es1[i]<- 0
poblacion$Es1<-Es1

Es2<-numeric(length(poblacion$Estrato))
for (i in 1:length(poblacion$Estrato)) if
(poblacion$Estrato[i]== "Estrato2") Es2[i] <- 1 else Es2[i]<- 0
poblacion$Es2<-Es2


###tamaños poblacionales en los estratos para tratamientos y controles
Nt=length(PT$Tk)
Nt1<-summary(PT$Estrato)[[1]]
Nt2<-summary(PT$Estrato)[[2]]
Nth=c(Nt1,Nt2)
Nc=length(PC$Tk)
Nc1<-summary(PC$Estrato)[[1]]
Nc2<-summary(PC$Estrato)[[2]]
Nch=c(Nc1,Nc2)
###tamanos muestrales
calmues2<-function(){
nt1=200
nt2=800
nth=c(nt1,nt2)
nc1=200
nc2=800
nch=c(nc1,nc2)
```

```
#### Estimacion en una muestra #########

###Supuestos conocimiento de totales
PT<-filter(poblacion, IT=="T")
PC<-filter(poblacion, IT=="C")
Tt1<-sum(PT$Tk*PT$Es1)
Tt2<-sum(PT$Tk*PT$Es2)
Tc1<-sum(PC$Tk*PC$Es1)
Tc2<-sum(PC$Tk*PC$Es2)
###Seleccion de muestras en los grupos tratamiento y control
samt <- S.STSI(PT$Estrato, Nth, nth)
Ptmues<- PT[samt, ]
#pesos ak en tratamientos
ak<-numeric(length(Ptmues$Estrato))
for (i in 1:length(Ptmues$Estrato)) if
(Ptmues$Estrato[i]== "Estrato1") ak[i] <-Nt1/nt1  else ak[i]<- Nt2/nt2
Ptmues$ak<-ak

samc <- S.STSI(PC$Estrato, Nch, nch)
Pcmues<- PC[samc, ]
##pesos ak en controles
ak<-numeric(length(Pcmues$Estrato))
for (i in 1:length(Pcmues$Estrato)) if
(Pcmues$Estrato[i]== "Estrato1") ak[i] <-Nc1/nc1  else ak[i]<- Nc2/nc2
Pcmues$ak<-ak

datasam<-union(Ptmues,Pcmues)

############## Modelos con pesos diseño MAS

mTk<-glm(Tk~Zk, weights= ak, data=datasam)
Tk.hat <- predict(mTk, type="response")
datasam$Tk.hat<-Tk.hat
mdIV<-ivreg(Yk~Tk|Zk,weights= ak, data=datasam)
```

```
#########calibracion
######Supuestos
Nt1
Nt2
Nc1
Nc2
Tt1
Tt2
Tc1
Tc2
#poblacion tratamiento y control en la muestra
Ptm<-filter(datasam, IT=="T")
Pcm<-filter(datasam, IT=="C")

#Estimacion de pesos calf para tratamientos
Ptm1<-filter(Ptm, Es1==1)
Ptm2<-filter(Ptm, Es2==1)
####Estimacion del total para la variable TK.hat por HT
#en tratamientos y controles en cada estrato
Ttht1=sum(Ptm1$ak*Ptm1$Tk)
Ttht2=sum(Ptm2$ak*Ptm2$Tk)

TZkt1=Ptm1$Zk*t(Ptm1$Tk)
yZkt1= Ptm1$Zk*Ptm1$Yk
wkt1=(Nt1/nt1)*(1+((Ttht1-Tt1)*((Nt1/nt1)*sum(TZkt1))^(-1))*Ptm1$Zk)
Ptm1$wk<-wkt1

TZkt2=Ptm2$Zk*t(Ptm2$Tk)
yZkt2= Ptm2$Zk*Ptm2$Yk
wkt2=(Nt2/nt2)*(1+((Tt2-Ttht2)*((Nt2/nt2)*sum(TZkt2))^(-1))*Ptm2$Zk)
Ptm2$wk<-wkt2
Ptm<-union(Ptm1,Ptm2)
####Controles
Pcm1<-filter(Pcm, Es1==1)
Pcm2<-filter(Pcm, Es2==1)
####Estimacion del total para la variable TK.hat por HT
```

```
#en tratamientos y controles en cada estrato
Tcht1=sum(Pcm1$ak*Pcm1$Tk)
Tcht2=sum(Pcm2$ak*Pcm2$Tk)

TZkc1=Pcm1$Zk*t(Pcm1$Tk)
yZkc1= Pcm1$Zk*Pcm1$Yk
wkc1=(Nc1/nc1)*(1+((Tcht1-Tc1)*((Nc1/nc1)*sum(TZkc1))^(-1))*Pcm1$Zk)
Pcm1$wk<-wkc1
TZkc2=Pcm2$Zk*t(Pcm2$Tk)
yZkc2= Pcm2$Zk*Pcm2$Yk
wkc2=(Nc2/nc2)*(1+((Tcht2-Tc2)*((Nc2/nc2)*sum(TZkc2))^(-1))*Pcm2$Zk)
Pcm2$wk<-wkc2
Pcm<-union(Pcm1,Pcm2)

data<-union(Ptm,Pcm)
n<-length(data$Yk)

###Forma matricial
xb0<-rep(1,n)
data$xb0<-xb0
xx=matrix(c(data$xb0,data$Tk),ncol=2,nrow=n)
Zz=matrix(c(data$xb0,data$Zk),ncol=2, nrow=n)
Pz=Zz%*%solve(t(Zz)%*%Zz)%*%t(Zz)


###regresion con MC2E
akd<-diag(x=data$ak,ncol=n,nrow=n)
bethtES=(solve(t(xx)%*%Pz%*%akd%*%xx))%*%t(xx)%*%Pz%*%akd%*%data$Yk

wkd<-diag(x=data$wk,ncol=n,nrow=n)
betcalfES=(solve(t(xx)%*%Pz%*%wkd%*%xx))%*%t(xx)%*%Pz%*%wkd%*%data$Yk

##betas calibrados y H-T
beta0hEs<-data.frame(bethtES[1])
beta1hEs<-data.frame(bethtES[2])
beta0cEs<-data.frame(betcalfES[1])
```

```
beta1cEs<-data.frame(betcalfES[2])

par<-as.numeric(c(beta0hEs,beta0cEs,beta1hEs,beta1cEs))
par

}
nsim=1000
simST2 <- replicate(nsim, calmues2())
betST0hEs<-simST2[1,1:nsim]
betST0cEs<-simST2[2,1:nsim]
betST1hEs<-simST2[3,1:nsim]
betST1cEs<-simST2[4,1:nsim]
mean(betST0hEs)
mean(betST0cEs)
mean(betST1hEs)
mean(betST1cEs)
############Sesgo relativo de los betas
Br0hes=(1/nsim)*sum(betST0hEs-B0/B0)####betas
Br1hes=(1/nsim)*sum(betST1hEs-B1/B1)
Br0es=(1/nsim)*sum(betST0cEs-B0/B0)####betas con MC2E
Br1es=(1/nsim)*sum(betST1cEs-B1/B1)
############Eficiencia relativa
ER0es=sum(betST0cEs-B0)^2/sum(betST0hEs-B0)^2
ER1es=sum(betST1cEs-B1)^2/sum(betST1hEs-B1)^2
```

B.5. Diseño πPT con T_k continua

```
##################################################
############## Modelo Poblacional ###############
set.seed(201606)
N <- 10000

v=rnorm(N,mean=2,sd=2)
Ek<-rnorm(N)
Zk<-rnorm(N,mean= 3, sd= 5)
```

```
G0=0.7
G1=0.01
IT<-rbinom(N,1,0.5)
Tk<-G0+G1*Zk+3*Ek+v
poblacion<-data.frame(Zk)###Variable instrumental
poblacion$Tk<-Tk #variable continua explicativa
poblacion$IT<-IT #variable indicadora de tratamiento
B0=2
B1=7
poblacion$Yk<-B0+B1*poblacion$Tk+Ek ##variable de interes
#####Modelos poblacionales
Modelo1=lm(Yk~Tk, data=poblacion)
ModeloTk=lm(Tk~Zk,  data=poblacion)
summary(ModeloTk)###condicion de relevancia
cor(Ek,Zk)  ###condicion de exogeneidad
Tk.h<-predict(ModeloTk, type="response")
poblacion$Tk.h<-Tk.h
###Modelo con funcion ivreg
IV<-ivreg(Yk~Tk|Zk, data = poblacion)
###Identificacion de la poblacion en tratamientos y controles
IT2<-numeric(length(poblacion$IT))
for (i in 1:length(poblacion$IT)) if
(poblacion$IT[i]== 1) IT2[i] <- "T" else IT2[i]<- "C"
poblacion$IT<-IT2

PT<-filter(poblacion, IT=="T")
PC<-filter(poblacion, IT=="C")
cpobla=sum(PC$Tk)
Tpobla=sum(PT$Tk)
poblacion<-union(PT,PC)
##############seleccion de la muestra
calpipt<-function(){
n=1000
res<-S.piPS(n, poblacion$Tk, e=runif(N))
sam<-res[,1]
Pik.s<-res[,2]
```

```
datasam<-poblacion[sam,]
datasam$Pik.s<-Pik.s
############ Modelos con pesos diseño piPT
datasam$ak<-1/Pik.s
modTk<-lm(Tk~Zk, weights= ak, data=datasam)

Tk.hat <- predict(modTk, type="response")
datasam$Tk.hat<-Tk.hat
mdIV<-ivreg(Yk~Tk|Zk,weights= ak, data=datasam)
####################calibracion
######Supuestos

Nt=length(PT$Yk)
Nc=length(PC$Yk)

#poblacion tratamiento y control en la muestra
Ptm<-filter(datasam, IT=="T")
Pcm<-filter(datasam, IT=="C")
nt=length(Ptm$Yk)
nc=length(Pcm$Yk)

####Estimacion del total para la variable TK.hat
##por HT en tratamientos y controles
Ttht=sum(Ptm$ak*Ptm$Tk)
Tcht=sum(Pcm$ak*Pcm$Tk)
##Estimacion de pesos calf para tratamientos
TZkt=Ptm$Zk*t(Ptm$Tk)
yZkt= Ptm$Zk*Ptm$Yk
wkt=Ptm$ak*(1+((Tpobla-Ttht)*((Ptm$ak)*sum(TZkt))^(-1))*Ptm$Zk)
Ptm$wk<-wkt

####Controles
TZkc=Pcm$Zk*t(Pcm$Tk)
yZkc= Pcm$Zk*Pcm$Yk
wkc=Pcm$ak*(1+((Tcht-cpobla)*((Pcm$ak)*sum(TZkc))^(-1))*Pcm$Zk)
Pcm$wk<-wkc
```

```
data<-union(Ptm,Pcm)
##################Estimacion de Bcalf
xb0<-rep(1,n)
data$xb0<-xb0
xx=matrix(c(data$xb0,data$Tk),ncol=2,nrow=n)
Zz=matrix(c(data$xb0,data$Zk),ncol=2, nrow=n)
Pz=Zz%*%solve(t(Zz)%*%Zz)%*%t(Zz)
################regresion con MC2E
akd<-diag(x=data$ak,ncol=n,nrow=n)
betht=(solve(t(xx)%*%Pz%*%akd%*%xx))%*%t(xx)%*%Pz%*%akd%*%data$Yk

wkd<-diag(x=data$wk,ncol=n,nrow=n)
betcalf=(solve(t(xx)%*%Pz%*%wkd%*%xx))%*%t(xx)%*%Pz%*%wkd%*%data$Yk

beta0h<-data.frame(betht[1])
beta1h<-data.frame(betht[2])
beta0c<-data.frame(betcalf[1])
beta1c<-data.frame(betcalf[2])

par<-as.numeric(c(beta0h,beta0c,beta1h,beta1c))
par
}

nsim=1000
simST <- replicate(nsim, calpipt())
betST0h<-simST[1,1:nsim]
betST0c<-simST[2,1:nsim]
betST1h<-simST[3,1:nsim]
betST1c<-simST[4,1:nsim]

mean(betST0h)
mean(betST0c)
mean(betST1h)
mean(betST1c)
##########Sesgo relativo de los betas
Br0hp=(1/nsim)*sum(betST0h-B0/B0)####betas con MC2E
```

```
Br1hp=(1/nsim)*sum(betST1h-B1/B1)
Br0p=(1/nsim)*sum(betST0c-B0/B0)####betas con MC2E
Br1p=(1/nsim)*sum(betST1c-B1/B1)
##############Eficiencia relativa
ER0p=sum(betST0c-B0)^2/sum(betST0h-B0)^2
ER1p=sum(betST1c-B1)^2/sum(betST1h-B1)^2
```

Apéndice C

Código en R de la aplicación en escenario real

C.1. Código en R

```
#####Modelos poblacionales
load("F:/tesis/BASES DANE/bastotal76.rda")
fix(bastotal76)

#####Modelos poblacionales
Mod1=lm(propun~ inpos + ESTU_EDAD + ESTU_AREA_RESIDE + FAMI_EDUCA_MADRE +
FAMI_PERSONAS_HOGAR +ESTU_TRABAJA + COLE_VALOR_PENSION, data=bastotal76)
summary(Mod1)

###Modelo con funcion ivreg
IV<-ivreg(propun ~inpos +ESTU_EDAD + ESTU_AREA_RESIDE + FAMI_EDUCA_MADRE +
FAMI_PERSONAS_HOGAR +ESTU_TRABAJA + COLE_VALOR_PENSION|TOT + IND +ESTU_EDA
 ESTU_AREA_RESIDE + FAMI_EDUCA_MADRE +FAMI_PERSONAS_HOGAR +ESTU_TRABAJA +
 COLE_VALOR_PENSION, data = bastotal76)
summary(IV)

####test de hausman de exogeneidad
```

```
cf_diff <- coef(IV)-coef(Mod1)
vc_diff<- vcov(IV)-vcov(Mod1)
xa_diff<-as.vector(t(cf_diff)%*% solve(vc_diff)%*% cf_diff)
pchisq(xa_diff, df=10, lower.tail=FALSE)


N<-length(bastotal76$COLE_COD_ICFES)

##########################################################################
##########################################################################
#################### Estimacion en una muestra ###########################

###tamaño muestral
calmuc<-function(){
  n=1000

  ######Seleccion de muestra
  sam<- S.SI(N, n)
  datasam<- bastotal76[sam, ]
  datasam$ak <- rep(N / n, rep = n)

  ############## Modelos con pesos diseño MAS

  modTk<-lm(propun ~ inpos + ESTU_EDAD + ESTU_AREA_RESIDE + FAMI_EDUCA_M
  +FAMI_PERSONAS_HOGAR +ESTU_TRABAJA + COLE_VALOR_PENSION, weights= ak,
  summary(modTk)

  mdIV<-ivreg(propun ~inpos +ESTU_EDAD + ESTU_AREA_RESIDE + FAMI_EDUCA_M
  FAMI_PERSONAS_HOGAR +ESTU_TRABAJA + COLE_VALOR_PENSION|TOT + IND +ESTU
  ESTU_AREA_RESIDE + FAMI_EDUCA_MADRE +FAMI_PERSONAS_HOGAR +ESTU_TRABAJA
  COLE_VALOR_PENSION,weights= ak, data=datasam)
  summary(mdIV)



  ################################calibracion
```

```
######Supuestos

Tpobla=sum(bastotal76$inpos)

####Estimacion del total para la variable TK.hat por HT en tratamientos


Ttht=sum(datasam$ak*datasam$inpos)

# Estimacion de pesos calf para tratamientos

TZkt=datasam$TOT*t(datasam$inpos)
yZkt= datasam$TOT*datasam$propun
wkt=(N/n)*(1+((abs(Tpobla-Ttht))*((N/n)*sum(TZkt))^(-1))*datasam$TOT)
datasam$wk<-wkt


##################Estimacion de Bcalf
xb0<-rep(1,n)
datasam$xb0<-xb0
xx=matrix(c(datasam$xb0,datasam$inpos, datasam$ESTU_EDAD, datasam$ESTU_A
datasam$FAMI_EDUCA_MADRE ,datasam$FAMI_PERSONAS_HOGAR,datasam$ESTU_TRABA
datasam$COLE_VALOR_PENSION),ncol=8, nrow=n)
Zz=matrix(c(datasam$xb0,datasam$TOT,datasam$IND, datasam$ESTU_EDAD,
datasam$ESTU_AREA_RESIDE, datasam$FAMI_EDUCA_MADRE ,datasam$FAMI_PERSONA
datasam$ESTU_TRABAJA, datasam$COLE_VALOR_PENSION),ncol=9, nrow=n)
Pz=Zz%*%solve(t(Zz)%*%Zz)%*%t(Zz)


###############regresion con MC2E
akd<-diag(x=datasam$ak,ncol=n,nrow=n)
bethtc=(solve(t(xx)%*%Pz%*%akd%*%xx))%*%t(xx)%*%Pz%*%akd%*%datasam$propu

wkd<-diag(x=datasam$wk,ncol=n,nrow=n)
betcalfc=(solve(t(xx)%*%Pz%*%wkd%*%xx))%*%t(xx)%*%Pz%*%wkd%*%datasam$pro
```

```
  ivewk<-ivreg(propun ~inpos +ESTU_EDAD + ESTU_AREA_RESIDE + FAMI_EDUCA_
  FAMI_PERSONAS_HOGAR +ESTU_TRABAJA + COLE_VALOR_PENSION|TOT + IND +ESTU
  ESTU_AREA_RESIDE + FAMI_EDUCA_MADRE +FAMI_PERSONAS_HOGAR +ESTU_TRABAJA
  COLE_VALOR_PENSION,weights= wk, data=datasam)
  summary(ivewk)

  beta0hc<-data.frame(bethtc[1])
  beta1hc<-data.frame(bethtc[2])
  beta2hc<-data.frame(bethtc[3])
  beta3hc<-data.frame(bethtc[4])
  beta4hc<-data.frame(bethtc[5])
  beta5hc<-data.frame(bethtc[6])
  beta6hc<-data.frame(bethtc[7])
  beta7hc<-data.frame(bethtc[8])
  beta0cc<-data.frame(betcalfc[1])
  beta1cc<-data.frame(betcalfc[2])
  beta2cc<-data.frame(betcalfc[3])
  beta3cc<-data.frame(betcalfc[4])
  beta4cc<-data.frame(betcalfc[5])
  beta5cc<-data.frame(betcalfc[6])
  beta6cc<-data.frame(betcalfc[7])
  beta7cc<-data.frame(betcalfc[8])


  par<-as.numeric(c(beta0hc,beta1hc,beta2hc,beta3hc,beta4hc,beta5hc,beta
  beta0cc,beta1cc,beta2cc,beta3cc,beta4cc,beta5cc,beta6cc,beta7cc))
  par
}
nsim=100
simSTc <- replicate(nsim, calmuc())
betST0hc<-simSTc[1,1:nsim]
betST1hc<-simSTc[2,1:nsim]
betST2hc<-simSTc[3,1:nsim]
betST3hc<-simSTc[4,1:nsim]
betST4hc<-simSTc[5,1:nsim]
betST5hc<-simSTc[6,1:nsim]
```

```
betST6hc<-simSTc[7,1:nsim]
betST7hc<-simSTc[8,1:nsim]
betST0cc<-simSTc[9,1:nsim]
betST1cc<-simSTc[10,1:nsim]
betST2cc<-simSTc[11,1:nsim]
betST3cc<-simSTc[12,1:nsim]
betST4cc<-simSTc[13,1:nsim]
betST5cc<-simSTc[14,1:nsim]
betST6cc<-simSTc[15,1:nsim]
betST7cc<-simSTc[16,1:nsim]


mean(betST0hc)
mean(betST1hc)
mean(betST2hc)
mean(betST3hc)
mean(betST4hc)
mean(betST5hc)
mean(betST6hc)
mean(betST7hc)
mean(betST0cc)
mean(betST1cc)
mean(betST2cc)
mean(betST3cc)
mean(betST4cc)
mean(betST5cc)
mean(betST6cc)
mean(betST7cc)
#########################Sesgo relativo de los betas
Br0hc=(1/nsim)*sum(betST0hc-47.176157 /47.176157 )####betas con HT
Br1hc=(1/nsim)*sum(betST1hc-0.258803 /0.258803 )
Br0c=(1/nsim)*sum(betST0cc-47.176157 /47.176157 )####betas con calf
Br1c=(1/nsim)*sum(betST1cc-0.258803 /0.258803 )
#########################Eficiencia relativa
ER0c=sum(betST0cc-47.176157)^2/sum(betST0hc-47.176157)^2
ER1c=sum(betST1cc-0.258803)^2/sum(betST1hc-0.258803)^2
```

Bibliografía

Victor M. Estevao and Carl-Erik Sarndal. (2000), A Functional Form Approach to calibration. *Journal of official Statistics. Vol.16*, (No.4.), pp. 379-399.

Yves Tillé. Teoria de Muestreo: Estimacíon de la varianza por linealización. Groupe de Statistique, Université de Neuchâtel. 2005.

Joshua D. Angrist and Jorn-Steffen Pischke. Mostly Harmless Econometrics: Instrumental Variables. March 2008.

H. Andrés Gutiérrez R. Estrategias de muestreo.Diseños de encuestas y estimacion de parámetros. Edición 1. Bogotá D.C.: Universidad Santo Tomas, 2009. ISBN: 978-958-631-608-8.

H. Andrés Gutiérrez R. & Viviana Natalia Rivera. (2012), Un acercamiento a la estimación de totales mediante la calibración de razones auxiliares en encuestas complejas. *Revista ib DANE*, *Vol 2.* (Num 2.)

H. Andrés Gutiérrez R. Modelos para estimar cambios brutos en encuestas rotativas: Pseudo-verosimilitud. Bogotá D.C., 2014, 178 h. Trabajo de grado(Doctor en ciencias- Estadística). Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Estadística.

Raquel Bernal & Ximena Peña. Guía Práctica para la Evaluación de Impacto. Edición 1. Bogotá D.C.: Universidad de los Andes,Facultad de Economía 2010. ISSN: 0120-3584.

Leonardo Bonilla M. y Luis A. Galvis Profecionalización docente y calida de la educación en colombia. 2012, 50 pag.

DNP (Departamento Nacional de Planeación). SINERGIA. Guías metodológiacas: Guía para la evaluación de políticas públicas. 174 pag.

Instituto Nacional de Estadística Uruguay. Impuestos IRPF (Categoría II). Diseños Muestrales: Ponderadores Calibrados. Anexo 1. 21 h.

Luis Francisco Rincón Suarez. Curso Básico de Modelos Lineales: Modelos de regresión lineal. Edición 1. Bogotá D.C.: Universidad Santo Tomas, 2009. pag 17. ISBN: 978-958-691-610-1.