Regresión lineal bajo diseños muestrales complejos: un enfoque aplicado

Víctor Morales Oñate; Bolívar Morales Oñate

Instituto de Estadística, Universidad de Valparaíso, Valparaíso, Chile / Facultad de Ingeniería en Ciencias y Alimentos, Universidad Técnica de Ambato, Ambato, Ecuador

victor.morales@uv.cl / be.morales@uta.edu.ec*

Resumen

Al utilizar una encuesta realizada bajo muestreo probabilístico, el investigador se enfrenta a un dilema en cuanto a inferencia, que puede ser basada en el modelo o en el diseño muestral. En este artículo se hace una revisión teórico-práctica de este dilema y se aplica ambos enfoques tanto a la estimación puntual como al ajuste de un modelo de regresión, evidenciando las diferencias en la Encuesta de Ingresos y Gastos de Hogares Urbanos 2002-2003 (EIGHU) configurada bajo un diseño muestral complejo.

Palabras clave: regresión, diseños complejos, muestreo.

Abstract

When using a survey conducted under probabilistic sampling, the researcher faces quandary with respect to inference, which can be based on the model or the sample design. In this article, a theoretical-practical review of this dilemma is carried out and both approaches are applied to a point estimate and to the adjustment of a regression model, demonstrating the differences in the Urban Household Income and Expenditure Survey 2002−2003 (EIGHU) which is configured under a complex sampling design.

Keywords: regression, complex survey, sampling.

Clasificador JEL: C13,C51,C83,D31

*Los autores agradecen a Liliana Roldan por su colaboración expedita.

1 Introducción

El uso de información proveniente de encuestas es la forma más usual en la que trabajan los Institutos de Estadística en América Latina. Sin embargo, cada país presenta diferencias metodológicas en algunos indicadores que se desprenden de las encuestas, a pesar de que éstos tratan de guardar cierta estandarización para procurar comparabilidad en las cifras a nivel internacional. Por otro lado, los usuarios de las encuestas en el sector público, sector privado y en la academia, a veces pueden dar a estas fuentes de información un tratamiento diferente de aquel para el cual se construyó el indicador.

Parte del tramado metodológico común que mantienen las encuestas en América Latina es que son realizadas mediante *muestreo probabilístico*. Por ejemplo, en Ecuador, la Encuesta Nacional de Empleo y Subempleo (ENEMDU) y la Encuesta Nacional de Ingresos y Gastos de los hogares urbanos y rurales (ENIGHUR) tienen un diseño de selección de la muestra que es probabilístico bietápico[1]. Este tipo de muestreo se alínea con lo que se conoce como muestreo *basado en el diseño*, lo cual genera cierta sutileza estadística al momento del tratamiento que se le da a estas fuentes de información, particularmente en lo que se refiere a inferencia.

Tres de los enfoques más usados para hacer inferencia son i) la teoría de inferencia estadística bajo el supuesto de poblaciones infinitas (teoría que se puede encontrar en cualquier libro de estadística matemática como en Rice (2006) o Bickel y Doksum (2015)), ii) inferencia basada en el diseño (Neyman, 1934) y iii) inferencia *basada en el modelo* (Matérn, 1960). Claramente, encuestas como la ENEMDU o la ENIGHUR se refieren a poblaciones finitas, y es claro que siguiendo los lineamientos de la sección 2.3, es viable usar el tipo de inferencia (i). En general, surge una pregunta de modo natural, ¿Qué tipo de inferencia usar? Desde el punto de vista teórico, se reconoce que los tres paradigmas tienen una base sólida, diferenciándose fundamentalmente en el objetivo propuesto (Schreuder *et al.*, 1993, pág. 205)[2]. Desde un enfoque práctico, sea el objetivo el cálculo de un indicador puntual o los coeficientes de un modelo de regresión, el valor de los parámetros estimados bajo los tres enfoques son los *mismos*, pero se diferencian en el valor de su respectiva varianza, de ahí su importancia en cuanto a inferencia poblacional. Un investigador debe tener muy presente estas diferencias al momento de interpretar sus estimaciones.

Basados en Gregoire (1998), el presente trabajo hace una revisión teórico para delimitar las diferencias entre la inferencia basada en el diseño y la inferencia basada en el modelo. Luego, mediante una aplicación en la Encuesta de Ingresos y Gastos de Hogares Urbanos 2002-2003 (EIGHU) del INEC, se muestra que optar por una u otra perspectiva tiene un impacto considerable al momento de hacer el cálculo de indicadores y en la estimación de modelos de regresión. Específicamente, la aplicación contrasta los resultados del modelo de *Las diferencias salariales entre el sector público y privado en el Ecuador* estimado en Carrillo (2004); tanto de la estimación de las variables del modelo a modo de indicadores, así como

[1] Para mayor detalle en cada una de los diseños muestrales refiérase a www.inec.gob.ec

[2] Ver secciones 2.1, 2.2 y 2.3

de la estimación de los coeficientes producto de la regresión lineal bajo diseños muestrales complejos.

2 Marco Teórico

Quizá una de las disputas más documentadas en la epistemología estadística es la Bayesiana vs Frecuentista. Ésta considera los parámetros del modelo como fijos y desconocidos, mientras aquella plantea que los parámetros son aleatorios (tienen una distribución de probabilidad asociada). Largas discusiones se han presentado en la literatura en cuanto a la lógica *más apropiada* a emplearse, pero actualmente parece haber un consenso: una u otra serán válidas en función del desafío de modelización específico que se aborde (Efron, 2005). En lo referente a la inferencia basada en el diseño o el modelo, parece haber un consenso similar, pero es imperativo que el investigador este consciente de los supuestos vinculados a ambos paradigmas. En términos generales, la diferencia entre la inferencia basada en el diseño (IBDI) y la inferencia basada en el modelo (IBMO), es que la primera considera a la población como fija y la segunda como si la población obedece a un proceso aleatorio. De ahí que resulte una cierta analogía en la querella filosófica Bayesiana-Frecuentista, IBMO-IBDI. Para un lector familiarizado con la estadística paramética y no paramétrica, puede resultar más preciso *alinear* a la IBMO con la primera la IBDI con la segunda.

La tabla 1 muestra los puntos más relevantes del devenir histórico de lo que hoy se conoce como *teoría de muestreo*, mismo que engloba a la IBDI y a la IBMO. El período comprendido entre 1662-1952 se caracteriza por tener a la IBDI como predominante, ya al final de intervalo es donde se empieza a cuestionar fuertemente esta perspectiva. Luego, entre 1952 y 1976 la discusión sobre la pertinencia de la IBDI o la IBMO está puesta sobre la mesa en diferentes congresos internacionales de estadística. Finalmente, con la aparición del libro *Model Assited Survey Sampling* de C.E. Särndal, desde 1977 se empieza a visualizar una reconciliación entre ambas posturas, definiendo sus alcances y convergencias. Sin embargo, una difusión aplicada de estos acuerdos ha permanecido en esferas académicas y es menester que se llegue a los usuarios más frecuentes: los institutos de estadística.

Tabla 1: Hitos históricos en la teoría de muestreo

Año	Hito
1662	Primera estimación mediante razonamiento estadístico (en el sentido actual) a partir de una muestra (Graunt, 1665)[3].
1901	Se demuestra empíricamente que, seleccionando muestras estratificadas, se obtienen mejores resultados en las estimaciones de medias y totales (Kiaer, 1901).

[3]La referencia se encuentra en (Galindo, 2007, pág. 7). Imágenes digitales del documento original se encuentra en (for the History of Science, 2017)

Tabla 1: (continuación)

Año	Hito
1906	Uso de aproximaciones de la distribución normal para la estimación de proporciones y propuesta de fórmula para estimación de varianza en muestreo estratificado (Bowley, 1906).
1924	Se crea una comisión de discusión del método representativo (ISI, 1924).
1926	Propuesta de métodos de selección representativos con probabilidades de inclusión iguales (Bowley, 1926).
1927	Publicación de tablas de números aleatorios (Tippett, 1927).
1927	Se publica el artículo considerado como uno de los pilares del muestreo como se conoce hoy en día. Libera el muestreo de las probabilidades de inclusión iguales. Introdujo en su artículo las ideas de eficiencia, asignación óptima, generalización del teorema de Markov, muestreo por conglomerados y presenta un caso donde el muestreo por conveniencia lleva a resultados equivocados (Neyman, 1934).
1939	Se introduce la técnica ANOVA para estimar la ganancia en eficiencia debida a la estratificación, se propone también la estimación de la varianza para muestras en dos etapas(Cochran, 1939).
1940	Se introduce el estimador de razón y se desarrolla la teoría de la estimación de totales y medias mediante modelos de regresión (Cochran, 1940).
1944	Se introduce la teoría de muestreo sistemático (Madow, 1944).
1952	Se completa el fundamento de la inferencia basada en el diseño. Se proporciona un marco de trabajo para la teoría de muestreo proporcional sin reemplazo (Horvitz, 1952).
1955	Pone en tela de juicio el concepto de eficiencia al que Neyman se refería; se prueba que, bajo la inferencia basada en el diseño de muestreo, no existe un estimador insesgado de varianza mínima (Godambe, 1955).
1960	Ejemplo pionero de inferencia basada en modelos. Trabajo realizado para estimar variabilidad espacial (Matérn, 1960).
1971	Hasta entonces se usaba la inferencia basada en el diseño, pero Richard Royall propone rotundamente abandonar este enfoque (Royall, 1971, pág. 422).
1976	Se define la estimación basada en modelos (Smith, 1976).
1977	Se sugiere que se debe buscar una manera para que los estimadores tengan sentido en ambas doctrinas (Godambe y Thompson, 1977).

Tabla 1: (continuación)

Año	Hito
1984	Se implementan las sugerencias de Godambe de 1977 tratando de calibrar el diseño muestral de modo que *funcione* en ambas doctrinas (Sarndal, 1984).
1992	Se publica *Model Assited Survey Sampling*, aquí la inferencia se basa en el diseño pero la estrategia de muestreo se complementa con un modelo para la estimación del parámetro de interés. (Sarndal, 1992).

Elaborada en base a (Gutierrez, 2009, pág. 415) y ampliada por los autores.

Una de las consecuencias inmediatas del no reconocimiento de las diferencias entre ambas doctrinas, es que se puede renunciar al uso de los datos de una encuesta debido a la falta de, por ejemplo, su diseño muestral específico o su factor de expansión. Es decir, contar con éstos elementos (diseño y factor de expansión) nos permite hacer IBDI e IBMO. Pero si no se precisa de esta información, aún se puede usar esa encuesta bajo el sólido fundamento de la IBMO. Además, la confusión entre ambos paradigmas puede resultar en inferencias sin validez alguna (Gregoire, 1998).

Se establece a continuación generalidades notacionales de teoría de muestreo cuando se tiene una población finita. Sea $\mathcal{U} = \{u_1, \ldots, u_k, \ldots, u_N\}$ una población finita de N elementos con etiquetas $k = 1 \ldots, N$. Y es la *variable de estudio* -cualitativa o cuantitativa- y Y_k denota el valor del k-ésimo elemento de la población $\mathcal{U}$. También se suele contar con un vector de información auxiliar X'_k de dimensión $p \times 1$. Así, el objetivo es la estimación de una función $g(T_y)$, donde los casos más usados son, $T_y = \sum_{k \in \mathcal{U}} Y_k$ para el total, $g(T_y) = T_y/N$ para la media y $g(T_y) = T_y/T_x = R$ para la razón. Sea Ω el conjunto de todas las muestras posibles y sea $p(\cdot)$ una función tal que $p(s)$ devuelve la probabilidad de seleccionar cualquier muestra s de la variable aleatoria S (la función $p(\cdot)$, también conocida como *diseño muestral*, determina la distribución de probabilidad de S). Sea I_k una variable aleatoria de inclusión muestral ($I_k = 1$ si se selecciona el k-ésimo elemento o $I_k = 0$ en caso contrario). La probabilidad de que un elemento k sea incluido en la muestra bajo un diseño $p(\cdot)$ es:

$$\pi_k = Prob(I_k = 1) = \sum_{S \in \Omega} I_k p(s) = \sum_{S \in \Omega_k} p(s) \tag{1}$$

donde $S \in \Omega_k$ denota que la suma es sobre todas las muestras s que contienen un k dado. Finalmente, $\nu = \sum_{k \in \mathcal{U}} I_k$ denota el número de elementos distintos en una muestra de tamaño n[4].

[4]Para más detalle y ejemplos véase (Sarndal, 1992, págs. 27-48) y (Gregoire, 1998).

2.1 Inferencia basada en el diseño

Bajo este arquetipo, la población es considerada fija y la muestra como una realización de un proceso estocástico. La inferencia se basa en la distribución de las estimaciones generadas por el diseño muestral (conocida como *distribución de referencia* (Fisher, 1956)). Es decir, Y_k no tiene ningún supuesto distribucional y la inferencia de las propiedades estadísticas de los estimadores se basa en la distribución de las estimaciones que resulta de todas las posibles muestras permisibles bajo el diseño muestral.

El estimador del total poblacional $T_y = \sum_{k\in\mathcal{U}} y_k$, también conocido como estimador de Horvitz-Thomson (HT), es uno de los más usados en la literatura:

$$\hat{T}_y = \sum_{k\in S} \frac{y_k}{\pi_k} = \sum_{k\in\mathcal{U}} \frac{y_k I_k}{\pi_k}. \tag{2}$$

Note que y_k es fijo, por tanto lo único aleatorio en $\hat{T}_y$ es el cómo opera I_k para que el elemento k sea incluido en la muestra. Se puede demostrar que $\hat{T}_y$ es un estimador insesgado de T_y y que su varianza es:

$$V\left[\hat{T}_y\right] = \sum_{k\in\mathcal{U}} y_k^2 \left(\frac{1-\pi_k}{\pi_k}\right) + \sum_{k\neq k'\in\mathcal{U}}\sum y_k y_{k'} \left(\frac{\pi_{kk'} - \pi_k\pi_{k'}}{\pi_k\pi_{k'}}\right) \tag{3}$$

donde $\pi_k = E\left[I_k\right] = E\left[I_k^2\right] = Prob[I_k = 1]$, y $\pi_{kk'} = Prob[I_k = 1, I_{k'} = 1]$ es la probabilidad de inclusión conjunta de k y k'.

De (3) se aprecia que la varianza de $\hat{T}_y$ es la varianza de las estimaciones de todas las muestras de Ω, esto es, depende de la distribución de referencia. Es decir que, en la IBDI, la varianza de un estimador cualquiera no es estadísticamente dependiente de la distribución de $y_k \in \mathcal{U}$.

En el caso de regresión lineal, suponga que cuenta con información de toda la población $\mathcal{U}$ y los parámetros de interés son los coeficientes $B_1, \ldots, B_p$ del modelo de regresión. Entonces, mediante el método de mínimos cuadrados ordinarios,

$$\mathbf{B_{p\times 1}} = \left(\mathbf{X_{p\times N}}\mathbf{X'_{p\times N}}\right)^{-1}\mathbf{X_{p\times N}}\mathbf{y_{N\times 1}}. \tag{4}$$

Por ejemplo, si $p = 2$, x_{2k} es el ingreso disponible y y_k son los ahorros del k-ésimo hogar ($k = 1, \ldots, N$), en el ajuste $y_k = B_1 + B_2 x_{2k}$ ($x_{1k} = 1$), B_2 representa el ahorro adicional generado por un dólar extra de ingreso disponible en la población. En el caso de tener una muestra s de un diseño muestral -por ejemplo, estratificado- deben tomarse en cuenta los pesos ($w_k = 1/\pi_k$) de cada estrato para evitar el sesgo en la estimación de los coeficientes (para más detalle ver (Cochran, 1977, págs. 189-203)). Es importante mencionar que para la estimación de los coeficientes no se ha hecho ningún supuesto distribucional de $y|x$. $\mathbf{B}$ sólo se considera como una característica que describe un aspecto de la población finita $\mathcal{U}$ que se desea estimar (Sarndal, 1992, pág. 190). Es decir, en la IBDI, el marco en el que se hace

inferencia es *exclusivamente* de la población en sí misma, ya sea que el parámetro de interés sea el total, la media, la razón o los coeficientes de una regresión. En resumen,

- Se prescinde de la idea de que la población ha sido aleatorizada, los y_k se consideran fijos y asociados a un elemento de la población finita $\mathcal{U}$.
- El diseño $p(\cdot)$ y un estimador específico T generan una distribución de estimaciones llamada distribución de referencia.
- La distribución de referencia induce las propiedades estadísticas de los esimadores.
- La inferencia se hace sobre el *ahora*, sobre el estado actual de la población $\mathcal{U}$.

2.2 Inferencia basada en el modelo

Así como a una muestra se la puede considerar como una *subpoblación* de $\mathcal{U}$, también existe el concepto de *superpoblación*. En la IBMO la población es considerada como una realización de un proceso aleatorio, un modelo ξ o *superpoblación*. Es decir, los valores $y_1, \ldots, y_N$ son realizaciones de las variables aleatorias $Y_1, \ldots, Y_N$ (Gregoire, 1998) donde éstas constituyen la superpoblación.

Sea $\hat{\theta}_s$ un estimador de θ y ξ el modelo asumido. En esta configuración la inferencia puede ser con respecto a un parámetro de la población $(g(T_y))$ o de la superpoblación (θ), tal que

$$E_\xi[(\hat{\theta}_s - \theta)^2 | s] \tag{5}$$

sea lo más pequeño posible. Es decir, se buscar minimizar el error cuadrático medio dado la muestra s.

Observación 1. Särndall propone un criterio *intermedio* a (5) de la forma $E_\xi E_{p(s)}[(\hat{\theta}_s - \theta)^2]$. Este criterio tomaría en cuenta tanto el diseño $p(s)$ como el modelo asumido ξ de modo que la inferencia sea circunscrita exclusivamente a la población finita $\mathcal{U}$ (Sarndal, 1992, pág.516). Sin embargo, (Gregoire, 1998, obs.4) aclara que esta distinción se desvanece, pues uno esperaría que asumiendo un modelo ξ apropiado y el tamaño de la muestra es suficientemente grande, la media y la varianza serían cercanas al de la población o la superpoblación según sea el caso.

Note que en (5) la selección de s es crucial. El diseño muestral no juega ningún papel en la inferencia cuando la distribución de referencia es establecida por el modelo asumido. Sin embargo, si el modelo falla es preferible el criterio de la observación 1.

Ejemplo 1. Considere el modelo de regresión a través del origen[5]

$$Y_k = \beta X_k + \sigma \epsilon_k \sqrt{X_k}, \epsilon_k \sim \mathcal{N}(0,1), cov(\epsilon_k, \epsilon_k') = 0$$

[5]Ejemplo tomado de (Gregoire, 1998)

Usando el método de mínimos cuadrados ordinarios para toda la población ($k = 1, \ldots, N$), se tiene

$$\hat{\beta} = T_y/T_x = R$$

Claramente, $\hat{\beta}$ es un estimador insesgado de β:

$$\begin{aligned}
E_m\left[\hat{\beta}\right] &= \frac{E_m\left[T_Y\right]}{T_X} = \frac{1}{T_X} E_m\left[\sum_{k\in\mathcal{U}} Y_k\right] \\
&= \frac{1}{T_X} E_m\left[\sum_{k\in\mathcal{U}} \beta X_k + \sigma\epsilon_k\sqrt{X_k}\right] \\
&= \frac{1}{T_X} \sum_{k\in\mathcal{U}} E_m\left[\beta X_k + \sigma\epsilon_k\sqrt{X_k}\right] \\
&= \frac{1}{T_X} \sum_{k\in\mathcal{U}} \left(\beta X_k + E_m\left[\sigma\epsilon_k\sqrt{X_k}\right]\right) \\
&= \frac{1}{T_X} \sum_{k\in\mathcal{U}} \left(\beta X_k + \sigma E_m\left[\epsilon_k\right]\sqrt{X_k}\right) \\
&= \frac{1}{T_X} \sum_{k\in\mathcal{U}} \beta X_k = \beta
\end{aligned}$$

donde E_m es la esperanza del modelo. Dado que $V_m[Y_k|X_k] = \sigma^2 X_k$, entonces

$$V\left[\hat{\beta}\right] = \frac{1}{T_X^2} \sum_{k\in\mathcal{U}} \sigma^2 X_k = \frac{\sigma^2}{T_X}$$

Así, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2/T_X)$.

Como se aprecia en el ejemplo 1, incluso después de la estimación, β y σ^2 siguen siendo parámetros desconocidos, de ahí la gran diferencia la con el paradigma de la inferencia basada en el diseño. Note también que la estimación de los coeficientes del modelo pudo ser por el método de máxima verosimilitud, lo cual no es factible bajo la visión de la IBDI. En resumen:

- La población se considera como la realización de un proceso estocástico.
- La distribución de probabilidad del modelo induce una distribución de referencia.
- Las propiedades de los estimadores dependen de la muestra y del modelo asumido.
- El diseño muestral es irrelevante para la inferencia, pero un muestreo probabilístico puede ayudar a disminuir errores de especificación del modelo.

2.3 ¿Cuál elegir?

Los libros clásicos de teoría de muestreo y artículos más recientes evitan dar reglas *fijas* para elegir una postura, pero todos tienen secciones de discusión para que el investigador pueda guiarse en su elección (Cochran, 1977; Lumley, 2011; Sarndal, 1992). Sin embargo, siguiendo a (Sarndal, 1992, pág. 515), es posible delimitar lo que el investigador *puede* hacer en función de la información que dispone y el objetivo que busca. En el caso de regresión lineal, y contando con un tamaño muestral grande, la tabla 2 resume las opciones del investigador.

Tabla 2: Tipos de inferencia para regresión lineal bajo diseños complejos

Parámetros de interés	**Tipo de inferencia**
A. Parámetros de la población finita $\mathcal{U}$	Inferencia basada en el diseño
B. Parámetros de la población finita $\mathcal{U}$	Inferencia basada en el modelo
C. Parámetros de la superpoblación ξ	Inferencia de teoría clásica de regresión lineal

Elaborada por autores en base a (Sarndal, 1992, pág. 515).

El primer caso (A) de la tabla 2 se refiere la sección 2.1, el segundo (B) a la observación 1 y el tercero (C) al ejemplo 1. Como se puede apreciar, tanto la observación como el ejemplo se ubican en el marco de la IBMO. Esto se debe a que para n grande, se pueden considerar *análogos*.

En una encuesta de muestreo probabilístico como la ENEMDU o la ENGHUR, el caso *óptimo* es contar con la variable de interés Y_k, el diseño muestral $p(\cdot)$, factor de expansión w_k y variables auxiliares X_k[6]. Para fines prácticos, los siguientes lineamientos indican cómo puede proceder el investigador si dispone de la información señalada:

Caso 1 (Y_k): Solo podría hacer estadística descriptiva de la subpoblación $k \in s$

Caso 2 (Y_k y w_k): Puede hacer estimación de los parámetros de la población finita $\mathcal{U}$. Debe tomar en cuenta que el cálculo de la varianza de los estimadores asume que el muestreo probabilístico fue realizado bajo muestreo aleatorio simple, de no ser éste el caso (como de hecho ocurre con la ENEMDU o la ENIGHUR), la inferencia se ve afectada.

Caso 3 (Y_k, w_k y $p(\cdot)$): En este caso se puede hacer estimación e inferencia sobre la población finita $\mathcal{U}$ bajo la lógica de la IBDI.

[6]De hecho, sería mejor aún contar con las variables con las que se construye los factores de expansión, pero este es no es el caso para el público en general.

Caso 4 (Y_k, w_k, $p(\cdot)$ y X_k): En este caso puede optar por cualquiera de las opciones de la tabla 2, esto es, IBDI o IBMO.

2.4 Discusión

Se han descrito los métodos de inferencia más utilizados en la literatura tradicional. Además se ha proporcionado una guía de elección entre IBDI e IBMO en función del instrumento disponible, de forma general y también acotado para el caso ecuatoriano. Pero ambas filosofías se presentan en muchas más aplicaciones. Actualmente, áreas de vanguardia en investigación estadística han adoptado la misma distinción en sus respectivas disciplinas y, por lo tanto, es imperativo que el investigador tenga presente la distinción original. Por ejemplo, en el campo de la estimación de áreas pequeñas (*SAE* por sus siglas en inglés), trabajos como Longford (2010) prefieren el uso de la inferencia basada en el diseño. Ahí se propone una combinación lineal convexa, aprovechando la similaridad espacial de los dominios, de modo que se utiliza información auxiliar (de censos o registros administrativos) y encuestas de muestreo probabilístico simultáneamente. Otros aportes como Rao y Molina (2015) presentan un repaso general del enfoque IBDI para SAE en los primeros capítulos de su libro. Pero principalmente su trabajo está basado en el enfoque IBMO, pasando por modelos lineales generalizados y modelos jerárquicos bayesianos. La estadística espacial también hereda la distinción IBDI e IBMO. Por ejemplo, de Gruijter J. *et al.* (2006) señala que la idoneidad de los dos enfoques se puede expresar en función de la resolución espacial en la que se trabaja. Si se trabaja con datos de áreas geográficas, se plantea a la IBDI como una mejor elección que la IBMO. Pero si se requiere mayor resolución, como la longitud y latitud de las estaciones en una red de monitoreo, el enfoque IBMO es más apropiado.

Por otro lado, en la misma dirección que la observación 1, a partir de la aparición de *Model Assited Survey Sampling* en 1992, se han generado varios aportes a la literatura en cuanto a la posibilidad de generar modelos *híbridos* que combinen ambos paradigmas. En el mismo espíritu que Gregoire (1998), pero con una recopilación bibliográfica más actualizada en lo respecta a modelos híbridos, Sterba (2009) hace un barrido de las posibles opciones. Un modelo se considera *híbrido* debido a que utiliza información del diseño muestral. Por ejemplo, el factor de expansión puede utilizarse como los pesos en la estimación de los parámetros del modelo de regresión lineal mediante el método de los mínimos cuadrados generalizados. Otra forma de obtener un modelo híbrido es la incorporación del diseño en la estimación de la varianza del estimador ($Var(\hat{\beta})$). Un efecto similar se obtiene en la sección 4.2 del presente trabajo, debido a que se corrigen los errores por el método de White. Dos de las ventajas más importantes de los modelos híbridos recogidas en Sterba (2009), son: i) posibilidad de inferencia y análisis causal en poblaciones finitas e infinitas y, ii) tienen la posibilidad de tomar en cuenta el error de medición. El segundo punto excluye enfáticamente al marco IBDI, pero en el primero se debe tener ciertas precauciones. Por ejemplo, dado que es un modelo híbrido, si se desea hacer un análisis causal, obviamente este modelo *hereda* el

problema de la especificación del enfoque IBMO. Además, para beneficiarse de las ventajas de un modelo híbrido, es necesaria una muestra grande. Al no ser este el caso, como se ejemplifica en la secciones 4.2 y 4.2, el investigador debe tener más precauciones al elegir el enfoque a usarse.

3 Metodología

Con el objeto de exponer las diferencias entre los paradigmas de la IBDI y la IBMO, se usa el modelo de *Las diferencias salariales entre el sector público y privado en el Ecuador* estimado en (Carrillo, 2004). En primer lugar, cada una de las variables que participan del modelo -tanto la dependiente como las independientes- son consideradas como variables objetivo. Es decir, se compara la estimación de las variables como si fuesen indicadores deseados por el investigador, específicamente, los casos 1, 2 y 3 de la sección anterior. La comparación de los resultados se hace en función del coeficiente de variación de las estimaciones.

Luego, en el caso del modelo de regresión, se estima los coeficientes de regresión según la lógica IBDI e IBMO. Para la IBDI se usa el caso A y para el IBDI se usa el caso C de la tabla 2. De hecho, el ajuste usado en (Carrillo, 2004) corresponde a C. Finalmente, la comparación de modelos se hará a través del p-valor de cada coeficiente en los diferentes escenarios, a través de su nivel de significancia.

3.1 Los datos

La fuente de los datos utilizados es la Encuesta de Ingresos y Gastos de Hogares Urbanos 2002-2003 (EIGHU) del INEC. La muestra tiene aproximadamente 11 mil hogares. De aquí se hace una selección de 2812 asalariados incluyendo variables de ingresos, género, edad, escolaridad, etnia, estado civil y si la persona trabaja o no en el sector público. De este conjunto de 2812 asalariados se presentan estadísticas descriptivas en la tabla 3 y también se realiza un modelo de regresión lineal para subconjuntos de la misma, 4 estimaciones en total. El subconjunto $n_1 = 2812$ es obtenido bajo los criterios a) ser asalariado del sector formal, b) trabajar por lo menos 30 horas a la semana, c) no pertenecer a una organización cuya actividad esté vinculada con la Agricultura, Extracción de Petróleo, Manufactura, y Generación Eléctrica, y d) individuos cuya actividad, según el código CIIU3, se incluya en las categorías: J Intermediación financiera, K Actividades Inmobiliarias, L Administración pública y defensa, y planes de seguridad social, M Enseñanza, N Servicios sociales y de salud y, O Otras actividades comunitarias. Luego, $n_2 = 304$ cumple con los criterios a)-c) y, e) Enseñanza secundaria (códigos: 8020-8022), $n_3 = 286$ se obtiene como unn subconjunto de n_2 al considerar los criterios a)-c) y, f) Enseñanza primaria (códigos: 8000-8010). Finalmente, $n_4 = 256$ cumple con los criterios a)-c), y g) Trabajadores hospitalarios (código: 8511).

Tabla 3: Estadística descriptiva de la muestra $n_1 = 2812$.

Variables	Promedio	Desviación Estándar	Mínimo	Máximo
Ingreso mensual	370.76	317.26	100.8	4603.0
Horas de trabajo semanales	47.77	13.20	31.0	110.0
Salario horario	2.06	1.81	0.3	22.2
Mujer	0.41	0.49	0.0	1.0
Edad	38.64	11.99	16.0	85.0
Años de educación	13.20	4.12	0.0	21.0
Postgrado	0.03	0.18	0.0	1.0
Blanco	0.09	0.29	0.0	1.0
Indígena, negro o mulato	0.03	0.18	0.0	1.0
Casado	0.51	0.50	0.0	1.0
Sector público	0.51	0.50	0.0	1.0
Número de observaciones				2812

Elaborada por: autores
Fuente: EIGHU (2002-2003).

3.2 El modelo

El modelo para establecer los determinantes del salario, como se presenta en (Carrillo, 2004), es una ecuación semi-logarítimica:

$$ln(w_i) = X_i\beta + \delta P_i + \epsilon_i \tag{6}$$

donde, w_i es el salario por hora, X_i son variables explicativas que determinan el nivel del salario, β es un vector de parámetros, P_i es una variable dicotómica igual a uno si el individuo trabaja en el sector público, δ es un coeficiente escalar, y ϵ_i es una variable aleatoria que incluye todos los otros factores que forman parte del salario y que no son explicados por X_i. δ refleja el diferencial salarial (en términos porcentuales) entre los sectores público y privado.

Es preciso mencionar que, de forma implícita, (6) es concebido bajo el paradigma IBMO debido a que considera que ϵ_i es una *variable aleatoria*. En la perspectiva de la IBDI, ϵ_i no es considerada como una variable aleatoria sino como un error de estimación de la ecuación. En la siguiente sección se explican más diferencias.

4 Resultados de la estimación

Las estimaciones puntuales suelen ser el principal objetivo de este tipo de encuestas, estos resultados se presentan en la sección 4.1. Asimismo, en la sección 4.2, se muestra los resultados de la estimación del modelo de regresión. En ambos casos se aprecia las diferencias de

considerar (o no) el diseño muestral de la encuesta.

4.1 Estimadores Puntuales

En la tabla 3 se tiene la estadística descriptiva de la muestra. Ahora se considera el ejercicio de estimación puntual de cada uno de los promedios de esas variables como si fuesen el objetivo *per se*. La tabla 4 muestra los coeficientes de variación asociados a las estimaciones, en negrita se resalta los valores máximos por fila y en cursiva los máximos por columna del coeficiente de variación asociado a la estimación de cada parámetro. Note que *Muestra*, *Factor* y *Diseño* se relacionan directamente con los casos 1, 2 y 3 de la sección 2.3, se realiza la estimación puntual de los indicadores teniendo presente los supuestos de cada caso.

Tabla 4: Coeficiente de variación de la estimación del promedio.

Variables	Tipo	$n_1 = 2812$	$n_2 = 304$	$n_3 = 286$	$n_4 = 256$
Ingreso mensual	Muestra	1.614	2.522	2.352	**3.155**
	Factor	1.631	2.536	2.397	**3.152**
	Diseño	*2.134*	*2.698*	*2.781*	***3.255***
Horas de trabajo semanales	Muestra	0.521	1.149	0.682	**1.447**
	Factor	0.523	1.196	0.711	**1.448**
	Diseño	*0.608*	***1.533***	*0.779*	*1.511*
Salario horario	Muestra	1.656	2.657	2.408	**3.401**
	Factor	1.674	2.652	2.464	**3.387**
	Diseño	*2.223*	*2.662*	*2.834*	***3.476***
Mujer	Muestra	2.269	***5.379***	3.917	*4.651*
	Factor	2.271	**5.310**	*3.937*	4.609
	Diseño	*2.366*	**5.017**	3.752	4.555
Edad	Muestra	0.585	1.521	1.570	**1.713**
	Factor	0.586	1.547	1.545	**1.725**
	Diseño	*0.624*	*1.654*	*1.626*	***1.849***
Años de educación	Muestra	0.588	1.179	1.158	**2.079**
	Factor	0.584	1.173	1.132	**2.032**
	Diseño	0.794	1.296	1.201	**2.229**
Postgrado	Muestra	10.373	31.150	***32.862***	27.075
	Factor	10.174	30.160	**30.854**	26.992
	Diseño	*11.923*	***38.681***	32.054	*30.077*
Blanco	Muestra	5.947	16.749	23.563	**24.254**

Tabla 4: (Continuación)

Variables	Tipo	$n_1 = 2812$	$n_2 = 304$	$n_3 = 286$	$n_4 = 256$
	Factor	5.949	16.612	**24.026**	23.124
	Diseño	*7.543*	*17.744*	***25.299***	*24.281*
Indígena, negro o mulato	Muestra	9.926	37.420	25.178	**40.422**
	Factor	9.827	36.549	24.547	***42.042***
	Diseño	*11.895*	***43.351***	*27.839*	41.984
Casado	Muestra	1.842	4.999	5.369	**5.975**
	Factor	1.845	*5.078*	5.458	**6.006**
	Diseño	*2.133*	5.074	*6.005*	***6.353***
Sector público	Muestra	1.836	**4.450**	3.852	4.415
	Factor	1.844	**4.503**	*3.825*	4.442
	Diseño	*2.136*	***4.963***	3.807	*4.493*

Elaborada por: autores
Fuente: EIGHU (2002-2003).

Consecuentemente con la teoría, se puede apreciar que -en general- el coeficiente de variación aumenta a medida que el tamaño de la muestra disminuye y que sus valores más altos se reflejan en las variables *Postrado*, *Blanco* e *Indígena, negro o mulato* debido a que su estimador es cercano a cero. En casi todos los casos se tiene que el coeficiente de variación es mayor para *Diseño*. Sin embargo, bajo esta configuración, ésta estimación es mejor que los casos de *Muestra* y *Factor* dado que éstas suponen que las variables tienen una distribución normal (claramente, variables como el ingreso no se corresponde con este tipo de distribución). Esto se refuerza en el hecho de que la diferencia del coeficiente de variación entre Diseño y las demás es menor cuando la muestra es más pequeña. Es decir, mientras menor sea el número de individuos en la muestra existe mayor heterogeneidad en la estimación, aún asumiendo una distribución normal.

La estimación de los indicadores es la misma para los casos de *Factor* y *Diseño* pero su error estándar es diferente. En consecuencia, su coeficiente de variación y su intervalo de confianza también serán diferentes (ver anexo A). Esto claramente puede ocasionar el uso inadecuado de una estimación al suponer indebidamente que la variable sigue una distribución normal. En particular, al tomar el cociente de los coeficientes de variación de la tabla 4 $CV_{n=256}/CV_{n=2812}$, en casi todos los casos ,excepto en *años de educación*, *Diseño* es menor que *Muestra* o *Factor*. Esto evidencia que a medida de que la muestra es más pequeña, el supuesto distribucional se hace más sensible en relación a *Diseño*. Por lo expuesto, a menos que el investigador esté seguro que la variable en cuestión sigue una distribución normal, el camino más seguro a tomar será la estimación puntual bajo la lógica IBDI.

4.2 Estimación del Modelo

Esta sección analiza el modelo de *Las diferencias salariales entre el sector público y privado en el Ecuador* (Carrillo, 2004) con el objeto de enfatizar las diferencias en el uso de la lógica IBDI e IBMO. La validez del modelo desde la teoría económica no es discutido, se lo toma como referencia en el sentido netamente estadístico.

La tabla 5 muestra la estimación del modelo (6) desde la configuración IBDI e IBMO[7]. En cuanto a la IBMO, note que los errores estándar fueron corregidos por el método de White. Esto es una práctica común ya que toma en cuenta la heterocedasticidad. De hecho, como se señala en (Huber, 1967) y en (Lumley, 2011), cuando los errores estándar de los coeficientes de una regresión lineal son estimados por métodos robustos, su valores son casi los mismos tanto en la lógica IBDI como en la IBMO. La diferencia radica en la forma en que se maneja la estratificación para el cálculo de los errores estándar, en la IBMO dependerá de los errores y en la IBDI del diseño muestral. Esto se verifica en los resultados obtenidos en la tabla 5.

No obstante, note que se han marcado con negrita los coeficientes de *blanco*, *casado* y *postgrado* en los ajustes 1, 2 y 4 respectivamente. Esto responde a que el uso e interpretación de estos coeficientes estará sujeto el nivel de significancia que esté dispuesto a aceptar el investigador. En el primero se aprecia que puede ser usado a un nivel de significancia del 1 % bajo la lógica IBMO pero no bajo la lógica IBDI. En *casado* sucede lo contrario, y en el último, *postgrado*, sólo puede ser usado si el investigador acepta un nivel de significancia del 10 % en IBMO y del 5 % en IBDI.

Otro aspecto que sobresale en la estimación, es el hecho de que los valores de los coeficientes difieren ligeramente. Apreciar esta diferencia como *relevante* dependerá de la aplicación específica que realice el investigador. Pero, en general, se aprecia que a medida que la muestra disminuye, las diferencias se acentúan levemente. Esto se debe a que todos los modelos estimados en la tabla 5 pueden ser considerados como de *n grande*. Por ejemplo, (Cochran, 1977, págs. 195-198) realiza un ejercicio donde la estimación de la media a través de la técnica de regresión lineal bajo la lógica IBDI es superior a la media puntual con $n = 256$.

En cuanto a la interpretación, tomamos el coeficiente que refleja las diferencias salariales δ, que es el coeficiente de la variable *Sector público*: *se evidencia que los salarios del sector público son en promedio* 17,4 % *más altos que los del sector privado.* La anterior interpretación se refiere al enfoque IBMO. En la IBDI sería el mismo texto pero con el valor de 17,7 %.

Ahora, si el texto es el mismo, ¿en qué radica la diferencia? En el caso de la IBMO, el investigador hace referencia al *fenómeno social*, descubre o establece algo análogo a lo que sería una *ley* en las ciencias naturales. De hecho, se puede apreciar que el modelo es muy parecido -pero diferente- a la configuración de un modelo de Mincer (Mincer, 1974). En

[7]Con la salvedad de que en la lógica IBMO no se considera el error como una variable aleatoria, tal como se ha explicado en el marco teórico. Es decir, la aleatoridad se da en la selección de la muestra.

el caso del modelo para determinar el modelo de *Las diferencias salariales entre el sector público y privado en el Ecuador*, adecuadamente el investigador menciona de forma explícita que trata de capturar el diferencial salarial apoyado en literatura de economía laboral, específicamente en (Bender, 1998; Disney y Gosling, 1998; Rees y Shah, 1995). Así se confirma la intencionalidad de capturar el fenómeno social mencionado anteriormente.

Tabla 5: Determinantes del salario en el Ecuador, regresión bajo lógicas IBMO e IBDI. Variable dependiente: logaritmo del salario por hora.

Variables	(1) Total muestra		(2) Educación secundaria		(3) Educación primaria		(4) Trabajadores hospitalarios	
	IBMO	IBDI	IBMO	IBDI	IBMO	IBDI	IBMO	IBDI
Constante	0.241***	0.233***	0.712***	0.712***	0.401*	0.399*	-0.071	-0.085
	(0.084)	(0.090)	(0.259)	(0.247)	(0.210)	(0.224)	(0.264)	(-0.085)
Mujer	-0.082***	-0.085***	-0.003	0.004	-0.084**	-0.09**	-0.077	-0.075
	(0.020)	(0.020)	(0.042)	(0.039)	(0.037)	(0.035)	(0.053)	(-0.075)
Edad	0.021***	0.021***	0.004	0.001	0.021**	0.021**	0.044***	0.045***
	(0.004)	(0.004)	(0.013)	(0.013)	(0.010)	(0.010)	(0.011)	(0.045)
Edadˆ2	-1.E-04**	-1.E-04**	1.E-04	1.E-04	-4.E-05	-4.E-05	-4.E-04***	-4.E-04***
	(5.E-05)	(5.E-05)	(1.E-04)	(1.E-04)	(1.E-04)	(1.E-04)	(1.E-04)	(1.E-04)
Años de Educación	0.068***	0.068***	0.050***	0.052***	0.042***	0.043***	0.05***	0.049***
	(0.003)	(0.003)	(0.009)	(0.009)	(0.006)	(0.007)	(0.008)	(0.049)
Postgrado	0.294***	0.307***	-0.006	-0.01	0.108	0.094	**0.216***	**0.236****
	(0.065)	(0.076)	(0.080)	(0.087)	(0.117)	(0.117)	(0.110)	(0.236)
Blanco	**0.119*****	**0.114****	0	-0.011	0.116	0.092	0.041	0.052
	(0.040)	(0.050)	(0.075)	(0.068)	(0.084)	(0.096)	(0.105)	(0.052)
Indígena, negro o mulato	-0.109**	-0.111**	-0.178	-0.233	-0.111	-0.124	0.042	0.02
	(0.049)	(0.050)	(0.141)	(0.180)	(0.095)	(0.092)	(0.130)	(0.020)
Casado	0.130***	0.134***	**0.108****	**0.115*****	0.011	-0.005	0.084*	0.084*
	(0.020)	(0.022)	(0.045)	(0.043)	(0.041)	(0.045)	(0.048)	(0.084)
Sector público	0.174***	0.177***	-0.036	-0.018	0.099	0.095	0.225***	0.212***
	(0.022)	(0.025)	(0.057)	(0.055)	(0.060)	(0.066)	(0.057)	(0.212)
R2	0.373	0.375	0.325	0.351	0.402	0.396	0.410	0.409
Número de observaciones		2812		304		286		256

Errores estándar en paréntesis, IBMO corregidos por el método de White.
*: significativo al 10 %, **: significativo al 5 %, ***: significativo al 1 %
Fuente: EIGHU (2002-2003)

Finalmente, en la IBDI la interpretación cambia. Al interpretar 17,7 % del ajuste IBDI, el investigador debe tener claro que lo que está haciendo es una estimación *estática*. Se trata

de una *fotografía* de la situación particular del Ecuador para el año en el que se realiza la encuesta. En este caso, se refiere a que en la EIGHU 2002-2003 (y sólo en este período), los salarios del sector público son, en promedio, 17,7 % más altos que los del sector privado. Es más, cada coeficiente será interpretado de la misma manera: de forma aislada y en referencia exclusiva al período de la encuesta. En consecuencia, las variables explicativas del modelo se toman como información auxiliar para mejorar el ajuste, pero no como determinantes de la diferencia salarial como tal.

5 Conclusiones

Una revisión bibliográfica detallada permitió establecer lineamientos en cuanto al uso que se les puede dar a las encuestas de muestreo probabilístico. La operativización de estas ideas permitirá al investigador tener más claros los supuestos teóricos y prácticos que acompañan a cada uno de los casos discutidos en el cuerpo de este artículo. En particular, las pautas aquí presentadas permiten hacer uso de encuestas aún cuando falte información -por ejemplo- del factor de expansión o del diseño muestral.

Para discernir en el uso de los enfoques IBDI o IBMO, el investigador ahora está en capacidad de responder adecuadamente a aquello que considera *aleatorio*. En la lógica IBDI se genera aleatoriedad en función del diseño muestral, que a su vez genera una distribución de referencia. Mientras que en la IBMO, lo aleatorio responde al modelo ξ asumido.

Otra diferencia clara es que en la IBDI no se hace ningún supuesto distribucional y en la IBMO si. Es decir que incluso después de la estimación, los parámetros de la superpoblación permanecen desconocidos en la IBMO. Pero en la IBDI, cuando se tiene un censo, no existe aleatoriedad. Esto implica que una vez calculado el estimador, el parámetro poblacional es conocido.

Se ha evidenciado las diferencias en cuanto a la estimación puntual. El coeficiente de variación bajo la lógica IBDI es mayor que en la IBMO. Sin embargo, al prescindir de un supuesto distribucional, en este caso es más confiable la IBDI que la IBMO a menos que el investigador esté seguro del supuesto distribucional.

La significancia de los coeficientes de regresión lineal bajo diseños muestrales complejos también ha sido explorada. Se ha mostrado que ésta puede variar en algunos casos y se agudizan las diferencias a medida que se trabaja con tamaños muestrales más pequeños. Por otro lado, a pesar de las coincidencias en las estimaciones cuando se usan ajustes robustos en la regresión, siguen existiendo diferencias en la interpretación de los resultados bajo los paradigmas IBDI e IBMO.

Referencias

Bender, K. A. (1998). The central government-private sector wage differential. *Journal of Economic Surveys*, 12(2):177–220.

Bickel, P. J. y Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics*, volumen 2. CRC Press.

Bowley, A. (1926). Measurement of the precision attained in sampling.(annex a to the report by jensen.) bulletin of the international statistical institute, 22. *Supplement to*, 54(1):1–62.

Bowley, A. L. (1906). Address to the economic science and statistics section of the british association for the advancement of sciences. *Journal of the Statistical Royal Society*, 69(3):540–558.

Carrillo, P. (2004). Las diferencias salariales entre el sector público y privado en el ecuador. *Cuestiones Económicas*, 20(2:3):165–174.

Cochran, W. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34(34):492–510.

Cochran, W. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain of total produce. *Journal of Agricultural Science*, 30(30):262–275.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons.

de Gruijter J., D.J., B., M.F.P., B., y Knotters, M. (2006). *Sampling for Natural Resource Monitoring*. Springer, New York.

Disney, R. y Gosling, A. (1998). Does it pay to work in the public sector? *Fiscal Studies*, 19(4):347–374.

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100(469):1–5.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Hafner Publishing Co.

for the History of Science, M. P. I. (2017). Echo cultural heritage online. `http://echo.mpiwg-berlin.mpg.de`. Accessed: 8 de enero 2017.

Galindo, E. (2007). *Estadística elemental moderna, conceptos básicos y aplicaciones*. Prociencia Editores.

Godambe, V. y Thompson, M. (1977). Robust near optimal estimation in survey practice. *IS Bulletin*, 47:129–146.

Godambe, V. P. (1955). A unified theory of sampling for the finite populations. *Journal of the Royal Statistical Society*, 17(B17):73–96.

Graunt, J. (1665). *Natural and Political Observations Made upon the Bills of Mortality*. The Royal Society, 3 edici

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research*, 28(10):1429–1447.

Gutierrez, H. A. (2009). *Estrategias de Muestreo. Diseno de encuestas y estimacion de parametros*. Universidad Santo Tomas, Bogota.

Horvitz, D. . T. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(47):663–685.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. 1(1):221–233.

ISI (1924). Instituto internacional de estadistica.

Kiaer, A. N. (1901). Sur les methodes representatives ou typologiques.

Longford, N. T. (2010). Small area estimation with spatial similarity. *Computational Statistics & Data Analysis*, 54(4):1151–1166.

Lumley, T. (2011). *Complex surveys: a guide to analysis using R*, volumen 565. John Wiley & Sons.

Madow, W. G. . M. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15(15):1–24.

Matérn, B. (1960). Spatial variation. *Medd. Statens Skogsforskingsintitu*, 49(5):100–35.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. National Bureau of Economic Research.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Rao, J. N. y Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Rees, H. y Shah, A. (1995). Public-private sector wage differential in the uk. *The Manchester School*, 63(1):52–68.

Rice, J. (2006). *Mathematical statistics and data analysis*. Nelson Education.

Royall, R. M. (1971). Linear regression models in finite populations sampling theory. *Foundations of statistical inference*, pp. 259–279.

Sarndal, C. Swensson, B. . W. J. (1992). *Model Assited Survey Sampling*. Springer.

Sarndal, C. E. . W. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11(11):164–156.

Schreuder, H. T., Wood, G. B., y Gregoire, T. G. (1993). *Sampling methods for multiresource forest inventory*. John Wiley & Sons.

Smith, T. M. F. (1976). The foundations of survey sampling: a review (with discussion). *Journal of the Royal Statistical Society*, 139(2):183–204.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate behavioral research*, 44(6):711–740.

Tippett, L. (1927). Random number tables. *Tracts for Computers*, (15).

ANEXO

A Anexo 1

Tabla 6: Estimaciones puntuales de las variables analizadas en las secciones 4.1 y 4.2. E. Sdt: Error estándar.

Variables	Tipo	Media	E. Sdt	Media	E. Sdt	Media	E. Sdt	Media	E. Sdt
Ingreso mensual	Muestra	370.756	5.983	294.531	7.427	279.555	6.575	328.263	10.357
	Factor	375.945	6.132	295.219	7.486	282.039	6.761	326.562	10.293
	Diseño	375.945	8.021	295.219	7.966	282.039	7.844	326.562	10.629
Horas de trabajo semanales	Muestra	47.772	0.249	41.974	0.482	40.402	0.276	44.348	0.642
	Factor	47.837	0.250	42.116	0.504	40.537	0.288	44.306	0.642
	Diseño	47.837	0.291	42.116	0.646	40.537	0.316	44.306	0.670
Salario horario	Muestra	2.059	0.034	1.797	0.048	1.748	0.042	1.926	0.065
	Factor	2.086	0.035	1.795	0.048	1.760	0.043	1.916	0.065
	Diseño	2.086	0.046	1.795	0.048	1.760	0.050	1.916	0.067
Mujer	Muestra	0.409	0.009	0.533	0.029	0.696	0.027	0.645	0.030
	Factor	0.408	0.009	0.539	0.029	0.694	0.027	0.649	0.030
	Diseño	0.408	0.010	0.539	0.027	0.694	0.026	0.649	0.030
Edad	Muestra	38.639	0.226	41.964	0.638	43.434	0.682	42.113	0.721
	Factor	38.736	0.227	41.989	0.650	43.390	0.670	42.155	0.727
	Diseño	38.736	0.242	41.989	0.695	43.390	0.705	42.155	0.780
Años de educación	Muestra	13.196	0.078	15.418	0.182	14.923	0.173	13.270	0.276
	Factor	13.261	0.077	15.475	0.182	15.038	0.170	13.349	0.271
	Diseño	13.261	0.105	15.475	0.201	15.038	0.181	13.349	0.297
Postgrado	Muestra	0.032	0.003	0.033	0.010	0.031	0.010	0.051	0.014

Tabla 6: (Continuación)

Variables	Tipo	Media	E. Sdt	Media	E. Sdt	Media	E. Sdt	Media	E. Sdt
	Factor	0.033	0.003	0.035	0.011	0.036	0.011	0.051	0.014
	Diseño	0.033	0.004	0.035	0.014	0.036	0.011	0.051	0.015
Blanco	Muestra	0.091	0.005	0.105	0.018	0.059	0.014	0.063	0.015
	Factor	0.091	0.005	0.107	0.018	0.057	0.014	0.068	0.016
	Diseño	0.091	0.007	0.107	0.019	0.057	0.014	0.068	0.017
Indígena, negro o mulato	Muestra	0.035	0.003	0.023	0.009	0.052	0.013	0.023	0.009
	Factor	0.036	0.003	0.024	0.009	0.055	0.014	0.022	0.009
	Diseño	0.036	0.004	0.024	0.010	0.055	0.015	0.022	0.009
Casado	Muestra	0.512	0.009	0.569	0.028	0.549	0.029	0.523	0.031
	Factor	0.511	0.009	0.561	0.029	0.541	0.030	0.521	0.031
	Diseño	0.511	0.011	0.561	0.028	0.541	0.032	0.521	0.033
Sector público	Muestra	0.514	0.009	0.625	0.028	0.703	0.027	0.668	0.029
	Factor	0.511	0.009	0.619	0.028	0.706	0.027	0.665	0.030
	Diseño	0.511	0.011	0.619	0.031	0.706	0.027	0.665	0.030
Número de observaciones			2812		304		286		256

Elaborada por: autores
Fuente: EIGHU (2002-2003)