Modelo de regresión *logit* multinomial bajo diferentes diseños de muestreo

Tesis para optar al grado de Magíster en Estadística

María Eugenia Sotelo Rico

Profesor guía:
Carlos Felipe Henríquez Roldán, PhD

Valparaíso, Noviembre de 2016

Agradecimientos

En primer lugar, debo mencionar que cursé el programa de Magíster en Estadística y realicé esta tesis gracias a la beca otorgada por la Agencia Chilena de Cooperación Internacional para el Desarrollo.

En segundo lugar, quiero agradecer a la Facultad de Química de la Universidad de la República (Uruguay) por concederme la licencia durante mi estadía en Valparaíso.

Por último, agradezco especialmente a mis compañeros de la Unidad Académica de Educación Química de la Facultad de Química, María Noel, Alejandro, Ivana y Shirley, por el apoyo académico y laboral.

Índice general

Capítulo 1

Introducción

En la actualidad, es muy fácil contactarse con las personas de forma virtual. Existen programas informáticos sin costo para los usuarios que permiten "diseñar" y "analizar" encuestas en línea[1], las redes sociales presentan módulos de encuestas [2] y numerosos programas de televisión, radio y portales de internet realizan preguntas a manera de sondeo publicando los resultados a medida que van recibiendo respuestas. En este contexto, se podría pensar que cualquiera es capaz de realizar y analizar encuestas y el trabajo de los estadísticos parecería no tener sentido en esta particular área de aplicación. Dicha situación, y el desconocimiento de la importancia del plan de muestreo requerido, permite que se realicen encuestas con muestras de conveniencia, creyendo que con ellas se obtendrá una gran cantidad de casos y que serán representativos de determinada población. Esto puede ser útil para tener un primer acercamiento al estado del arte de lo que se quiere investigar pero puede llevar a conclusiones erróneas, debido a que siempre existirá el sesgo de selección; por lo cual, sería más conveniente tener una muestra de tamaño reducido pero obtenida bajo un diseño de muestreo correcto.

En el mismo sentido, es sencillo disponer de microdatos de encuestas, tanto en Chile como en Uruguay. En ambos países existen leyes promulgadas recientemente que permiten el acceso a la información pública (Ley N^o 20.285 de 2008, Chile y Ley N^o 18.381 de 2008, Uruguay). Si bien esto representa un logro, muchas veces se procesa la información creyendo que se conoce la teoría que hay detrás de la metodología de planificación y recolección de los datos y el investigador trabaja como si los datos hubiesen sido capturados utilizando un muestreo aleatorio simple; ya que, no incorpora el plan de muestreo planificado, que la mayoría de las veces no es un muestreo aleatorio simple.

[1] Formularios de Google, SurveyMonkey, LimeSurvey, entre otros

[2] Módulo "Haga preguntas. Obtenga respuestas" de Facebook y aplicación "Pollowers" de Twitter

También se hace necesario mencionar que existen trabajos que utilizan datos obtenidos por encuestas, que presentan el diseño de muestreo y lo utilizan en sus análisis.

Tomando como base las apreciaciones antes descritas, este trabajo pretende investigar la importancia de la incorporación de un plan de muestreo adecuado, tanto en la captura como en el análisis de los datos, cuando se quiere llevar a cabo una encuesta.

Como aplicación se utilizaron los datos del VII Censo de Estudiantes de Grado de la Facultad de Química de la Universidad de la República (Udelar) de Uruguay realizado entre setiembre y noviembre de 2012. A través del modelo *logit* multinomial se estudió la situación del estudiante y se estimó la probabilidad de que al 31 de Diciembre de 2015 un estudiante (que ha sido censado) egrese, abandone sus estudios o se mantenga activo como estudiante. Se trabajó bajo el supuesto de que a esa fecha los estudiantes censados deberían haber egresado, por cuanto las carreras en la Facultad de Química de la Udelar tienen una duración de cinco años, con excepción de la generación 2012[3]. El modelo propuesto se ajustó en la población y bajo diferentes diseños de muestreo, y se compararon las estimaciones de estos diseños con respecto a los parámetros poblacionales.

En lo que resta del capítulo introductorio se presentan la hipótesis, la pregunta de investigación, los objetivos y por último, se explica la situación de los estudiantes en la Facultad de Química de la Udelar de Uruguay, con el fin de contextualizar los datos que se utilizaron en este trabajo. En el capítulo dos se describe el marco teórico del diseño muestral y del modelo de regresión *logit* multinomial. El capítulo tres y cuatro presentan la metodología utilizada y los resultados obtenidos, respectivamente. Por último, en el capítulo cinco, se plantean las conclusiones a las que se arribaron en esta investigación.

1.1. Hipótesis y pregunta de investigación

En este trabajo se plantea la siguiente hipótesis: si se considera un plan de muestreo complejo las estimaciones obtenidas en el modelo *logit* multinomial son más eficientes, consistentes y con errores estándares más pequeños, que si se considera un muestreo aleatorio simple.

La pregunta de investigación es ¿afecta el ajuste del modelo y por lo tanto las estimaciones puntuales y sus correspondientes errores estándares el plan de muestreo utilizado?

[3] Al 31 de diciembre de 2015 los estudiantes de la generación 2012 solamente pueden pertenecer a alguna de dos categorías: estudiante activo o desertor.

1.2. Objetivos

Objetivo general

El objetivo general de esta investigación es analizar las estimaciones del modelo *logit* multinomial bajo diferentes diseños de muestreo y compararlas con los parámetros poblacionales.

Objetivos específicos

En función del objetivo general, se proponen los objetivos específicos que se plantean a continuación:

(i) Establecer el tamaño de muestra mínimo necesario para especificar el modelo.

(ii) Investigar el ajuste del modelo bajo los diferentes planes de muestreo y comparar las estimaciones con los parámetros de la población.

(iii) Estudiar cómo varía el error estándar de las estimaciones según el diseño de muestreo propuesto.

(iv) Evaluar el test de bondad de ajuste para regresión *logit* multinomial propuesto por Fagerland, Hosmer y Bofin (2008).

1.3. Situación del estudiante en la Facultad de Química, Universidad de la República, Uruguay

En el año 2000 la Facultad de Química de la Udelar de Uruguay implementó un nuevo Plan de Estudios (Universidad de la República, 1999a), con el propósito de brindar a los egresados una apropiada formación y una información suficiente, atenuando algunos rasgos enciclopedistas característicos de los planes anteriores.

El Plan de Estudios 2000 (o Plan 2000)[4] está alineado con las tendencias actuales en educación superior, constituyéndose como una propuesta curricular centrada en el aprendizaje y en el trabajo del estudiante. Se ofrecen carreras de grado articuladas, lo que permite el tránsito horizontal estudiantil, y mediante un sistema de créditos y una oferta de asignaturas electivas se dota de flexibilidad a la estructura curricular.

[4]Ambas denominaciones son las utilizadas en los documentos oficiales de la Udelar

1.3. Situación del estudiante en la Facultad de Química, Universidad de la República, Uruguay

En el marco del Plan de Estudios 2000 la Facultad de Química de la Udelar otorga tres títulos profesionales, Químico Farmacéutico, Bioquímico Clínico y Químico, y un título académico de Licenciado en Química (Universidad de la República, 2003a). También imparte dos carreras profesionales compartidas con otras facultades, las cuales están articuladas con el Plan 2000: Ingeniería Química, compartida con la Facultad de Ingeniería (Universidad de la República, 1999b), e Ingeniería de Alimentos, compartida con las Facultades de Agronomía, Ingeniería y Veterinaria (Universidad de la República, 2003b). Las cinco carreras profesionales tienen una duración teórica de cinco años y requieren de 420-450 créditos para la titulación. La carrera de Licenciado en Química dura cuatro años, requiriéndose de 320 créditos para la obtención del título.

Dada la flexibilidad curricular y movilidad horizontal existente en el Plan 2000, un estudiante puede simultáneamente ser egresado de una o más carreras, ser estudiante activo de otra(s) carrera(s) y desertor de otra(s). Tanto la condición de estudiante activo como la condición de estudiante desertor pueden modificarse en el tiempo, y por lo tanto las definiciones requieren de una referencia temporal. En general, ambas condiciones se definen abarcando un período de dos años:

Estudiante activo: un estudiante es activo al 31 de diciembre del año X si realizó al menos una actividad académica en el período comprendido entre el 1 de enero del año $X-1$ y el 31 de diciembre del año X, y esta actividad no fue con el fin de egresar. Por ejemplo, un estudiante es activo al 31 de diciembre de 2016 si no ha egresado hasta esta fecha y si realizó al menos una actividad académica entre el 1 de enero del 2015 y el 31 de diciembre de 2016.

Estudiante desertor: un estudiante es desertor al 31 de diciembre del año X si su última actividad académica fue registrada en un año anterior al año $X-1$ y esta no fue con el fin de egresar. Por ejemplo, un estudiante es desertor al 31 de diciembre de 2016 si no ha egresado hasta esta fecha y su última actividad académica fue registrada en 2014 o antes.

Esta compleja situación se debe a que no existen impedimentos en la Udelar para que un estudiante retome sus estudios luego de haberlos abandonado. Esta Universidad es gratuita, no tiene examen de ingreso y no cuenta con restricciones a nivel de actividad académica del estudiante para continuar en el sistema (no hay plazos para aprobar los niveles, no hay límite de repeticiones, no hay límite de exámenes) (Boado, 2005).

Se entiende por actividad académica a la inscripción a curso o a examen, independientemente de su resultado. No están consideradas las inscripciones que fueron anuladas en tiempo y forma (desistimiento de cursos o exámenes).

Debido a que el Plan de Estudios 2000 permite la movilidad horizontal entre las carreras, algunos indicadores educativos tradicionalmente empleados (por ejemplo, tasa de abandono y porcentaje de egresos) no pueden aplicarse directamente y algunos requieren de una redefinición. Este tránsito horizontal habilita a que un estudiante pueda egresar de una carrera diferente de aquella por la cual ingresó a la Facultad, por lo tanto el tiempo de egreso medido desde la inscripción a la segunda carrera podría resultar en una subestimación del tiempo real de estudios, puesto que el estudiante puede haber revalidado asignaturas cursadas con anterioridad a la inscripción en la segunda carrera. Por otro lado, el egreso medido desde el ingreso a la Facultad podría resultar en una sobreestimación del tiempo real de estudios, puesto que el tiempo transcurrido entre el ingreso y la inscripción a la segunda carrera no necesariamente habrá de sumarse en todos los casos.

Capítulo 2

Marco teórico

Este capítulo consta de dos secciones. En la primera, se presenta brevemente la teoría que hay detrás de un diseño muestral y luego se describen los tipos de muestreos probabilísticos más comunes; y en la segunda, se exponen las etapas que se deben considerar cuando se trabaja con un modelo de regresión *logit* multinomial.

2.1. Diseño muestral

Los programas de análisis estadísticos realizan el procesamiento de datos que el investigador les indique, sin solicitar en ningún momento el plan de muestreo[1]. Por defecto los paquetes estadísticos están programados considerando que el plan de muestreo es un muestreo aleatorio simple. Si al analizar datos provenientes de una muestra no se tiene en cuenta el diseño de muestreo subyacente, las estimaciones obtenidas estarán sesgadas y por lo tanto llevarán a conclusiones erróneas.

El diseño muestral abarca dos componentes: el diseño de muestreo y el proceso de estimación. El plan de muestreo es la metodología utilizada para seleccionar la muestra de una población, y el proceso de estimación está constituido por los algoritmos o fórmulas utilizados para obtener las estimaciones en la muestra de los parámetros de la población (Levy y Lemeshow, 2008). Wolter (2007) resume la combinación de ambos componentes en la Tabla 2.1 y manifiesta que gran parte de la teoría de las encuestas por muestreo involucra el caso *a*, donde se proponen estimadores lineales (media o proporción, por ejemplo) bajo planes de muestreo aleatorio simple, pero que frecuentemente los estudios por encuestas presentan una mayor complejidad como en los casos *b*, *c* y *d*.

[1]En este trabajo el diseño muestral puede ser mencionado indistintamente como diseño de muestreo, plan de muestreo o esquema de muestreo

Tabla 2.1: Diseño muestral propuesto y tipo de estimadores a utilizar

Estimadores	Diseño	
	Simple	Complejo
Lineales	a	b
No lineales	c	d

Fuente: adaptado de Wolter (2007).

El autor mencionado plantea que una encuesta realizada con muestreo complejo incluye cinco dimensiones: (i) el grado de complejidad del diseño, (ii) el grado de complejidad de las estimaciones, (iii) múltiples características o variables de interés, (iv) el uso descriptivo y analítico de los datos de la encuesta y (v) la escala o tamaño de la encuesta. Los puntos (i) y (ii) fueron mencionados en el párrafo anterior. Con respecto a (iii), Wolter (2007) señala que en la actualidad en una encuesta -mediante muestreo complejo- se consideran diez o cien características de interés, a diferencia de lo que se hacía en el pasado donde solamente se tenía una variable de interés. La dimensión (iv) concierne la idea de que en una encuesta simple el objetivo es describir varias características de la población objetivo, mientras que en el muestreo complejo se pueden incluir objetivos analíticos como la construcción de modelos o las pruebas de hipótesis. Finalmente, la escala de la encuesta (dimensión (v)) es importante al momento de clasificar a una encuesta como simple o compleja. Una encuesta compleja implica cientos, si no miles, de individuos encuestados, además de un gran equipo para el levantamiento de datos.

La elección del diseño de muestreo debe ser un esfuerzo colaborativo entre el estadístico que diseñe la encuesta, las personas que participan en la ejecución de la encuesta y aquellas personas que utilizarán los datos relevados. Los usuarios de los datos especifican qué variables se deben medir, cuáles son las estimaciones requeridas, qué nivel de fiabilidad y validez se necesitan en las estimaciones, y cuáles son las restricciones en términos de plazos y costos. Los participantes de la ejecución presentan los costos de personal, tiempo y materiales necesarios, así como también la viabilidad de los procedimientos de muestreo y medición. Habiendo recibido esta información, el profesional de estadística puede proponer el diseño de muestreo teniendo en cuenta las especificaciones con el menor costo posible (Levy y Lemeshow, 2008).

Existen dos maneras de extraer una muestra de una población: probabilística o no probabilística. Ejemplos del primer caso son el muestreo aleatorio simple, el estratificado, por conglomerados y el sistemático. Mientras que ejemplos no probabilísticos son el muestreo por cuotas, por bola de nieve, de conveniencia (también llamado casual), de juicio experto, y *respondent driven sampling*. La principal ventaja del muestreo probabilístico es

la posibilidad de calcular estimadores insesgados de los parámetros poblacionales a partir de los datos de la muestra, así como también su error estándar. Por el contrario, el no probabilístico no cuenta con esta característica y por lo tanto no se puede evaluar la calidad de las estimaciones.

Kalton (1983) menciona que el muestreo no probabilístico es ampliamente utilizado por razones de costo o conveniencia. Sin embargo, se puede mencionar que: (i) si se utiliza un muestra de conveniencia, como es el caso de las encuestas en televisión o en portales de internet, se debe tener en cuenta que a partir de los resultados no se pueden hacer inferencias de la población en general; y (ii) en el muestreo de juicio de expertos se solicita a un experto en el tema que se quiere investigar que seleccione una muestra "representativa" de la población (por ejemplo, un investigador en educación elige un grupo de escuelas de una ciudad), pero una muestra que para un determinado experto es "representativa" puede no serlo para otro, lo cual hará que la muestra tenga un sesgo que no se puede calcular. Dicho lo anterior, en este trabajo se aplicarán las técnicas de muestreo probabilístico, las cuales serán presentadas en el apartado 2.1.1.

A continuación se definen brevemente conceptos importantes que se deben conocer a la hora de especificar un diseño de muestreo: población, marco muestral y muestra. Existe una amplia bibliografía en la que se definen estos conceptos (Zeng, 2011; Heeringa, West y Berlung, 2010; Levy y Lemeshow, 2008; Lehtonen y Pahkinen, 2004; Lohr, 2009; Särndal, Swensson y Wretman, 1992):

Población (universo o población objetivo) es un conjunto finito de elementos (o unidades) del que se quiere obtener información y estimar los parámetros. Independientemente del tamaño, en teoría cada elemento de la población podría ser contado en un censo o ser seleccionado en la muestra para ser encuestado. En general, el listado con la población objetivo no se encuentra disponible y se debe recurrir al marco muestral.

Marco muestral es el listado que identifica y permite acceder a los elementos de la población, y tiene la propiedad de que cada unidad perteneciente a este listado tiene alguna chance de ser seleccionada en el momento de extraer la muestra. Los elementos de la población que están incluidos en el marco constituyen lo que se denomina población marco (conjunto U de tamaño N). El marco muestral incluye información auxiliar que permite realizar técnicas de muestreo complejo (por ejemplo, estratificación) y técnicas especiales de estimación (por ejemplo, estimación de razón).

Muestra es un subconjunto de U y existen muchas muestras diferentes que pueden ser extraídas. El conjunto de M muestras posibles de tamaño n ($n < N$) de U es denotado por S, $S = \{s_1, s_2, ..., s_M\}$. La muestra real es denotada por $s = 1, ..., i, ..., n$

y es una de las posibles muestras de S. Para seleccionar una muestra de U se utiliza un esquema de muestreo específico.

2.1.1. Tipos de muestreo

Existen diferentes formas de seleccionar muestras aleatorias a la hora de llevar a cabo una encuesta. El método más conocido y sencillo es el muestreo aleatorio simple (MAS), pero también existen otros, como el estratificado, por conglomerados, el sistemático y la combinación de al menos dos de éstos. La característica que tienen en común es que se conoce o se puede conocer la probabilidad de inclusión de cada elemento antes de llevar a cabo la selección de la muestra. Estos elementos pueden ser personas, hogares, países, empresas, instituciones educativas, u otros. En el caso del MAS cada elemento tiene la misma probabilidad de selección, pero en el resto de los métodos la probabilidad de inclusión es diferente entre las unidades de la población.

Para cualquier elemento $i \in U$ la probabilidad de inclusión es π_i. Cuando se introduce más de una técnica para seleccionar una muestra en un diseño de muestreo, π_i es la probabilidad de inclusión de primer orden.

Muestreo aletorio simple

El muestreo aleatorio simple se define como un esquema de muestreo donde cualquiera de los posibles subconjuntos de n elementos distintos de la población U tiene la misma probabilidad de ser seleccionado, lo cual implica que todas las unidades de la población tienen la misma probabilidad de ser incluidas en la muestra (Kalton, 1983), $\pi_i = \pi = n/N,\ \forall\, i \in U$.

La muestra es seleccionada sin utilizar información auxiliar de la población. Por esta razón, el MAS provee una referencia para evaluar la ganancia de la información auxiliar en un esquema de muestreo más complejo o en una mejora de la estimación (Lehtonen y Pahkinen, 2004).

Muestreo estratificado

El muestreo estratificado consiste en la clasificación de la población en subpoblaciones (estratos) basada en información auxiliar, para luego seleccionar muestras independientes desde cada estrato. Lehtonen y Pahkinen (2004) plantean que a menudo la información auxiliar corresponde a características inherentes a la población, como pueden ser las demográficas o socioeconómicas. El tamaño muestral en cada estrato es controlado por el muestrista y no por el azar (Kalton, 1983), y puede ser mediante asignación óptima, asignación proporcional o asignación *ad hoc*, lo cual produce que en este esquema los errores

estándares sean más pequeños que en el MAS (Heeringa *et al.*, 2010).

Heeringa *et al.* (2010) mencionan que la estratificación puede ser utilizada para seleccionar elementos o conglomerados de elementos. Los estratos no se solapan, son homogéneos y los define la persona encargada del esquema de muestreo antes de seleccionar la muestra.

En este tipo de muestreo la probabilidad de selección de los elementos de los estratos dependerá de la técnica seleccionada. El caso más sencillo es cuando se realiza un MAS; de esta manera, la probabilidad de selección del $i - ésimo$ elemento, $i = 1, ..., h$, del estrato $h - ésimo$ es $\pi_{hi} = n_h/N_h$.

De acuerdo a lo propuesto por Heeringa *et al.* (2010) un diseño de muestreo estratificado implica los siguientes cuatro pasos en la extracción de la muestra y el análisis de los datos:

1. Se forman los estratos de N_h elementos o conglomerados $(h = 1, ..., H)$.

2. De manera independiente en cada uno de los estratos se seleccionan muestras de a_h conglomerados o $a_h = n_h$ elementos.

3. Con los casos de la muestra se calculan las estimaciones de los parámetros de interés separadamente en cada estrato y luego se ponderan y combinan para estimar el total de la población.

4. Se calculan las varianzas muestrales de las estimaciones separadamente para cada estrato y luego se ponderan y combinan para estimar la varianza muestral de la estimación del total poblacional.

Debido a que en el muestreo estratificado se extraen muestras independientes en cada uno de los $h = 1, ..., H$ estratos, cualquier varianza atribuible a las diferencias entre estratos es eliminada por la varianza de la estimación. Por consiguiente, se forman estratos internamente homogéneos (*within*) y externamente heterogéneos (*between*), obteniendo una variación intra estratos pequeña (Heeringa *et al.*, 2010 y Lehtonen y Pahkinen, 2004).

Muestreo por conglomerados

En el muestreo por conglomerados, al igual que en el muestreo estratificado, la población se encuentra dividida en grupos, pero la diferencia radica en que aquí se selecciona una muestra de las subpoblaciones y todos los elementos pertenecientes a estas subpoblaciones o una muestra aleatoria de éstos quedan incluidos en la muestra.

Este tipo de muestreo es empleado por varias razones: (i) en encuestas de hogares reduce los costos asociados a traslado debido a que los conglomerados se definen principalmente a partir de áreas geográficas; (ii) las unidades a ser seleccionadas no siempre son identificables en los marcos muestrales, pero pueden ser asociadas a un conglomerado. A menudo el marco muestral contiene los conglomerados, lo cual implica seleccionar una muestra de estos y luego es el encuestador quien selecciona las unidades a entrevistar; (iii) una o más etapas de la muestra son deliberadamente agrupadas para permitir estimaciones de modelos multinivel y componentes de varianzas en variables de interés (Heeringa *et al.*, 2010; Levy y Lemeshow, 2008 y Särndal *et al.*, 1992).

El mecanismo de selección es el siguiente:

1. La población U es particionada en M conglomerados, $U_I = \{U_1, U_2, ..., U_M\}$.
2. Se toma una muestra s de la población de conglomerados U_I bajo un diseño de muestro elegido por el investigador.
3. Se observan todos los elementos de U que pertenezcan a los conglomerados seleccionados o todos los elementos de U que fueron extraídos en una muestra dentro de cada conglomerado seleccionado.

Si bien este tipo de técnica reduce los costos y simplifica la logística de un trabajo con encuestas, se debe tener en cuenta que en la mayoría de los casos los errores estándar obtenidos con muestreo por conglomerados son mayores que los obtenidos con un MAS de igual tamaño. Esto se debe a que las observaciones que pertenecen a un mismo conglomerado son muy similares. También implica la estimación de una medida de homogeneidad denominada correlación intra-conglomerado.

De acuerdo a Zeng (2011), el muestreo por conglomerados puede ser clasificado en una etapa o en múltiples etapas e incluso en muestreo con probabilidades de inclusión iguales o con probabilidades de inclusión desiguales. En el muestreo en una etapa con probabilidades iguales los conglomerados son seleccionados mediante MAS. Una vez que el grupo es seleccionado aleatoriamente todas las unidades pertenecientes a él son incluidas en la muestra. En este caso el conglomerado es denominado unidad primaria de muestreo (UPM). La probabilidad de inclusión del $i-ésimo$ conglomerado es $\pi_i = m/M$, m tamaño de la muestra de conglomerados. Levy y Lemeshow (2008) demuestran que, en ciertas ocasiones, trabajar con probabilidades iguales puede llevar a seleccionar muestras que no son factibles de implementar y a que las estimaciones lineales tengan errores estándares más altos que cuando se trabaja con probabilidades desiguales. Estos problemas se presentan cuando existe una variabilidad considerable en la UPM con respecto a la cantidad de observaciones, lo cual es frecuente en la práctica.

Para el caso de muestreo en una etapa con probabilidades de selección desiguales, los conglomerados pueden ser seleccionados con probabilidad proporcional al tamaño (cantidad de elementos que lo componen): $\pi_i = N_i/N$, N_i es la cantidad de unidades que forman el conglomerado i.

En el muestreo en dos etapas las UPM son seleccionadas en una primera etapa de muestreo, y luego se extrae una submuestra de elementos en cada UPM seleccionada. Las unidades de la submuestra son denominadas unidades secundarias de muestreo (USM). Suponiendo un muestreo de conglomerados con probabilidad proporcional al tamaño, con π_i probabilidad de que el conglomerado i sea seleccionado, la probabilidad de inclusión de la $j-ésima$ unidad del conglomerado i es $\pi_{ij} = \pi_i n_i/N_i$.

2.1.2. Factores de expansión

En los planes de muestreo complejo es usual que las probabilidades de inclusión de las unidades sean diferentes entre sí, por lo tanto se hace necesario el uso de factores de expansión (también llamados ponderadores o pesos) para obtener estimadores insesgados y consistentes (Heeringa *et al.*, 2010 y Lehtonen y Pahkinen, 2004). La observación muestral ponderada es el número de elementos de la población que representa esa observación y se define como el inverso de la probabilidad de inclusión, $w_i = 1/\pi_i$, lo cual conlleva a que la esperanza $E\left(\sum_{i=1}^{n} w_i\right) = N$.

De acuerdo a las probabilidades de inclusión definidas en el apartado anterior los factores de expansión en cada tipo de muestreo son:

- MAS: $w_i = N/n$
- Estratificado: $w_{hj} = N_H/n_h$ es el factor de expansión de la unidad $j-ésima$ del estrato $h-ésimo$
- Conglomerado en una etapa: $w_i = M_i/m_i$ para el caso de probabilidades de inclusión iguales y $w_i = N/N_i$ para el caso de probabilidades de inclusión desiguales proporcionales al tamaño.

Según Zeng (2011), Heeringa *et al.* (2010) y Kalton (1983) las observaciones finalmente se ponderan por un factor de expansión final ($w_{final,i}$) con el objetivo de mejorar la calidad las estimaciones. Este ponderador es el producto del factor de la selección de la muestra ($w_{sel,i}$), un factor por no respuesta ($w_{nr,i}$) y un factor de post-estratificación ($w_{ps,i}$):

$$w_{final,i} = w_{sel,i} \times w_{nr,i} \times w_{ps,i}$$

El cálculo de los factores de expansión en muestreo complejo implica multiplicar las probabilidades de inclusión de cada etapa del muestreo y luego tomar el inverso de este

producto.

Debido a la no respuesta, en las encuestas se relevan efectivamente r observaciones de las n originales pertenecientes a la muestra ($r < n$). De acuerdo a Zeng (2011) uno de los métodos comúnmente utilizados para ajustar el $w_{sel,i}$ es el denominado ajuste de clase de ponderación, en el cual la muestra seleccionada se separa en varios grupos de clases de ponderación a partir de variables conocidas para todas las unidades de la muestra, tanto para las que respondieron como para las que no. Suponiendo una muestra s de tamaño n, la cual es dividida en K clases de ponderación de acuerdo a variables conocidas para todos los elementos de s, R_k es el conjunto de todos los respondentes en la $k-ésima$ clase, T_k es el conjunto de todos los elementos de la clase k, $k = 1, ..., K$, entonces el factor por no respuesta en la $k-ésima$ clase es:

$$\alpha_k = \frac{\sum_{i \in T_k} w_i}{\sum_{i \in R_k} w_i}$$

De esta manera el ponderador para la observación $i-ésima$ ajustado por no respuesta es $w_{nr,i} = w_{sel,i} \times \alpha_k$.

El ajuste por post-estratificación se realiza luego de que los datos son relevados y las observaciones se clasifican en grupos que son tratados como estratos (Zeng, 2011). Suponiendo que los elementos de la muestra son clasificados en G post-estratos, $g = 1, ..., G$, N_g es la población del post-estrato g, Q_g es el conjunto de elementos seleccionados en la muestra del $g-$ésimo post-estrato, entonces el factor de post-estratificación en cada post-estrato es

$$\delta_g = \frac{N_g}{\sum_{i \in Q_g} w_i}$$

Por lo tanto, el factor de expansión para la observación $i-ésima$ ajustado por post-estratificación es $w_{ps,i} = w_{sel,i}\delta_g$.

2.1.3. Efecto de diseño

El efecto de diseño es una medida de eficiencia de un plan de muestreo determinado en relación al MAS. Se define como el cociente entre la varianza del estimador bajo un diseño de muestreo y la varianza del estimador bajo MAS con la misma cantidad de observaciones:

$$deff_{PM}(\hat{\theta}) = \frac{V_{PM}(\hat{\theta})}{V_{MAS}(\hat{\theta})}$$

con $\hat{\theta}$ estimador del parámetro θ, $V_{PM}(\hat{\theta})$ varianza del estimador bajo un plan de muestreo determinado y $V_{MAS}(\hat{\theta})$ varianza del estimador bajo MAS.

Lohr (2009) manifiesta que el efecto de diseño provee una medida de la ganancia o pérdida de precisión por el uso de un diseño de muestreo complejo en lugar de un diseño aleatorio simple. Asimismo, dado que en un estudio por encuesta se puede obtener más de una estimación, se obtendrán diferentes efectos de diseño para cada una de ellas (Por ejemplo, en un modelo de regresión se obtendrán diferentes *deff* para cada uno de los $\hat{\beta}$).

Así, si se consideran varios planes de muestreo se obtendrán *deff* para cada uno de ellos. A modo de ejemplo se puede mencionar que, bajo muestreo estratificado con asignación proporcional, el efecto de diseño será menor que uno. Es decir, la estratificación es un plan de muestreo más preciso que el MAS. Por el contrario, el *deff* de trabajar con muestreo por conglomerados en una etapa es mayor que uno, lo cual conlleva en una pérdida de precisión con respecto al MAS.

El efecto de diseño se utiliza frecuentemente para estimar el tamaño de muestra necesario para llevar a cabo una encuesta. Si se conoce el *deff* para una encuesta similar, alcanza con estimar solamente el tamaño necesario bajo MAS y luego multiplicarlo por el *deff* para obtener la cantidad de observaciones requeridas bajo un diseño complejo (Lohr, 2009).

2.2. Modelo de regresión *logit* multinomial

El modelo *logit* multinomial es una extensión de los modelos de regresión para respuestas binarias a variables respuesta con tres o más categorías, y es la técnica de modelización apropiada para variables con categorías de respuesta nominal (Heeringa *et al.*, 2010).

Hosmer, Lemeshow y Sturdivant (2013) plantean que el modelo de regresión logística multinomial es utilizado para trabajar con una variable respuesta nominal de tres o más categorías y el objetivo es estimar la probabilidad de seleccionar cada una de las categorías, así como también estimar los *odds* en función de las covariables y expresar los resultados en términos de *odds ratio*.

A continuación se describen las etapas que implica la construcción del modelo *logit* multinomial: especificación, estimación, evaluación e interpretación.

2.2.1. Especificación del modelo

Utilizando como referencia a Hosmer *et al.* (2013), se presenta la especificación del modelo para el caso en que la variable respuesta acepta tres valores posibles, pero el desa-

rrollo puede ser extendido cuando la variable respuesta tiene más de tres categorías.

El modelo considera que las categorías de la variable respuesta, Y, son codificadas como 0, 1 y 2. Así, como en la regresión logística binaria, Y es reparametrizada en términos del *logit* de $Y = 1$ versus $Y = 0$, en este caso se necesitan dos funciones *logit*. El investigador es quien decide qué categoría toma como base o de referencia. Según Heeringa *et al.* (2010) la elección de la categoría base no afecta el ajuste del modelo ni los test de significación de los parámetros de los predictores, pero sí se debe tener en cuenta para la interpretación de las estimaciones.

De esta manera, si se considera un modelo con $Y = 0$ como categoría de referencia, p covariables y un término constante (el vector de covariables $\mathbf{x}$ tendrá $p+1$ columnas, con $x_0 = 1$), los *logits* son

$$\begin{aligned} g_1(\mathbf{x}) &= ln\left[\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] \\ &= \beta_{10} + \beta_{11}\, x_1 + \beta_{12}\, x_2 + ... + \beta_{1p}\, x_p \\ &= \mathbf{x}^\top \boldsymbol{\beta_1} \end{aligned} \tag{2.1}$$

y

$$\begin{aligned} g_2(\mathbf{x}) &= ln\left[\frac{P(Y=2|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] \\ &= \beta_{20} + \beta_{21}\, x_1 + \beta_{22}\, x_2 + ... + \beta_{2p}\, x_p \\ &= \mathbf{x}^\top \boldsymbol{\beta_2} \end{aligned} \tag{2.2}$$

Así, las probabilidades condicionales de cada respuesta dadas las p covariables son

$$P(Y=0|\mathbf{x}) = \frac{1}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}} \quad , \tag{2.3}$$

$$P(Y=1|\mathbf{x}) = \frac{e^{g_1(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}} \quad \text{y} \tag{2.4}$$

$$P(Y=2|\mathbf{x}) = \frac{e^{g_2(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}} \quad . \tag{2.5}$$

Si $\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$, $j = 0, 1, 2$, cada probabilidad es una función del vector de $2(p+1)$ parámetros $\boldsymbol{\beta}^\top = (\boldsymbol{\beta_1}^\top, \boldsymbol{\beta_2}^\top)$. De forma general,

$$\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x}) = \frac{e^{g_j(\mathbf{x})}}{\sum_{k=0}^{2} e^{g_k(\mathbf{x})}} \tag{2.6}$$

Al igual que en el modelo de regresión lineal, en este tipo de modelo se debe tener en cuenta la parsimonia. De acuerdo a Heeringa *et al.* (2010) el número de parámetros a estimar es $(j-1) \times (p+1)$, entonces para asegurar la eficiencia en las estimaciones y la precisión de la interpretación, la especificación final del modelo debe intentar minimizar la cantidad de predictores que no son significativos o que están altamente correlacionados con otras variables.

2.2.2. Estimación

Con el fin de hacer más simple la estimación, Hosmer *et al.* (2013) presentan el desarrollo para una variable dependiente con tres categorías, pero al igual que en la especificación, la estimación se puede extender a cualquier caso en que existan más de tres opciones para la variable respuesta.

El primer paso en la estimación es definir tres variables binarias (Y_0, Y_1 y Y_2) en la construcción de la función de verosimilitud, que indican a cuál de las categorías de respuesta -0,1,2- pertenece la observación:

$$\begin{aligned} &Y_0 = 1, \quad Y_1 = 0 \quad \text{y} \quad Y_2 = 0 \qquad \text{si} \quad y = 0 \\ &Y_0 = 0, \quad Y_1 = 1 \quad \text{y} \quad Y_2 = 0 \qquad \text{si} \quad y = 1 \\ &Y_0 = 0, \quad Y_1 = 0 \quad \text{y} \quad Y_2 = 1 \qquad \text{si} \quad y = 2 \end{aligned}$$

Sin importar el valor que tome Y, se cumple que $\sum_{j=0}^{2} Y_j = 1$. Así, la función de verosimilitud para una muestra de n observaciones independientes es

$$l(\beta) = \prod_{i=1}^{n} [\pi_0(\mathbf{x}_i)^{y_{0i}} \pi_1(\mathbf{x}_i)^{y_{1i}} \pi_2(\mathbf{x}_i)^{y_{2i}}] \tag{2.7}$$

Aplicando logaritmo y utilizando que $\sum y_{ij} = 1$ para cada i, la log-verosimilitud es

$$L(\beta) = \sum_{i=1}^{n} y_{1i} g_1(\mathbf{x}_i) + y_{2i} g_2(\mathbf{x}_i) - ln\left(1 + e^{g_1(\mathbf{x}_i)} + e^{g_2(\mathbf{x}_i)}\right) \tag{2.8}$$

Las ecuaciones de verosimilitud se resuelven tomando las primeras derivadas parciales de $L(\beta)$ con respecto a cada uno de los $2(p+1)$ parámetros desconocidos:

$$\frac{\partial L(\beta)}{\partial \beta_{jk}} = \sum_{i=1}^{n} \mathbf{x}_{ki}(y_{ji} - \pi_{ji}) \tag{2.9}$$

con $\pi_{ji} = \pi_j(x_i)$, $j = 1, 2$ y $k = 0, 1, 2, ..., p$ y $x_{0i} = 1$ para cada observación.

Debido a que las ecuaciones anteriores no son lineales, el estimador máximo verosímil $\hat{\beta}$ se obtiene igualando las ecuaciones a 0 y encontrando la solución mediante métodos iterativos.

Con respecto a la varianza, la estimación de la matriz de covarianzas de $\hat{\beta}$ se obtiene a partir de la matriz de segundas derivadas parciales, cuyos elementos son de la forma:

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{jk'}} = -\sum_{i=1}^{n} \mathbf{x}_{k'i}\, \mathbf{x}_{ki}\, \pi_{ji}\, (1 - \pi_{ji}) \tag{2.10}$$

y

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{j'k'}} = -\sum_{i=1}^{n} \mathbf{x}_{k'i}\, \mathbf{x}_{ki}\, \pi_{ji}\, \pi_{j'i} \tag{2.11}$$

para j y $j' = 1, 2$, y k y $k' = 0, 1, 2, ..., p$.

La estimación de la matriz de covarianzas del estimador máximo verosímil es el inverso de la matriz de información observada, $\hat{\mathbf{I}}(\hat{\beta})$. Esta matriz es de $2(p+1) \times 2(p+1)$ y sus elementos son los valores negativos de las ecuaciones 2.10 y 2.11 evaluadas en $\hat{\beta}$:

$$\widehat{V(\hat{\beta})} = \left[\hat{\mathbf{I}}(\hat{\beta})\right]^{-1}. \tag{2.12}$$

El estimador de la matriz de información puede representarse de la siguiente manera: sea $\mathbf{X}$ la matriz de $n \times (p+1)$ que contiene las covariables; sea $\mathbf{V_j}$ una matriz diagonal de $n \times n$ cuyos elementos son $\hat{\pi}_{ji}(1 - \hat{\pi}_{ji})$ para $j = 1, 2$ e $i = 1, 2, ..., n$; y sea $\mathbf{V_3}$ una matriz diagonal de $n \times n$ cuyos elementos son $\hat{\pi}_{1i}\hat{\pi}_{2i}$, entonces el estimador de la matriz de información se puede expresar como

$$\hat{\mathbf{I}}(\hat{\beta}) = \begin{bmatrix} \hat{\mathbf{I}}(\hat{\beta})_{11} & \hat{\mathbf{I}}(\hat{\beta})_{12} \\ \hat{\mathbf{I}}(\hat{\beta})_{21} & \hat{\mathbf{I}}(\hat{\beta})_{22} \end{bmatrix} \tag{2.13}$$

con $\quad \hat{\mathbf{I}}(\hat{\beta})_{11} = (\mathbf{X'V_1X}), \quad \hat{\mathbf{I}}(\hat{\beta})_{22} = (\mathbf{X'V_2X}) \quad$ y $\quad \hat{\mathbf{I}}(\hat{\beta})_{12} = \hat{\mathbf{I}}(\hat{\beta})_{21} = (\mathbf{X'V_3X})$.

Estimación bajo muestreo complejo

De acuerdo a Heeringa *et al.* (2010) cuando se estima un modelo de regresión *logit* multinomial con datos obtenidos bajo un diseño de muestreo complejo se utiliza una aproximación de la función de verosimilitud que incorpora los pesos muestrales, denominada pseudo-verosimilitud:

$$L(\beta) = \prod_{i=1}^{n} \left\{ \prod_{j=0}^{2} \pi_j(\mathbf{x}_i)^{ji} \right\}^{w_i} \tag{2.14}$$

con w_i factor de expansión de la i-ésima observación de la muestra.

La maximización implica la aplicación del algoritmo de Newton-Raphson para resolver las ecuaciones, asumiendo que el diseño de muestreo considera estratos indexados por h y conglomerados indexados por m:

$$\frac{\partial L(\beta)}{\partial \beta_{jk}} = \sum_h \sum_m \sum_i w_{hmi}\, \mathbf{x}_{hmki}\, (y_{hmji} - \pi_{ji}) \tag{2.15}$$

En el caso de la estimación de la varianza y covarianzas las segundas derivadas parciales son de la forma:

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{jk'}} = -\sum_h \sum_m \sum_i \mathbf{x}_{k'i}\, \mathbf{x}_{ki}\, w_{hmi}\, \pi_{ji}\, (1 - \pi_{ji}) \tag{2.16}$$

y

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{j'k'}} = -\sum_h \sum_m \sum_i \mathbf{x}_{k'i}\, \mathbf{x}_{ki}\, w_{hmi}\, \pi_{ji}\, \pi_{j'i} \quad . \tag{2.17}$$

2.2.3. Evaluación

Según Heeringa *et al.* (2010) la etapa de evaluación del modelo comienza con el test de Wald asociado a los parámetros del modelo. Con $(j-1) \times (p+1)$ parámetros estimados, el número de test de hipótesis posibles es casi ilimitado. De todas formas, en la práctica, tanto la prueba t para parámetros simples como el test de Wald para parámetros múltiples se utilizan para evaluar la significación del efecto de las covariables en los *logits* individuales, $H_0 : \beta_{jp} = 0$, o en todos los *logits* estimados $H_0 : \beta_{2p} = ... = \beta_{jp} = 0$. El estadístico de Wald se define como $W_{jp} = \hat{\beta}_{jp} / \hat{s}(\hat{\beta}_{jp})$, y bajo la hipótesis de que el coeficiente es cero sigue una distribución normal estándar.

Sin embargo, el test de Wald es un indicador preliminar de la importancia de la variable en el modelo, ya que por la cantidad de grados de libertad de las variables en este tipo de modelo se debe utilizar el test de razón de verosimilitud. En general, este test para la significación de los coeficientes de una variable tiene los grados de libertad igual al número de categorías de respuesta menos uno por los grados de libertad de la variable en cada *logit* (Hosmer *et al.*, 2013). Es decir, que si la variable respuesta tiene j categorías y la

covariable tiene c categorías los grados de libertad del test son $(j-1)\times(c-1)$.

Fagerland, Hosmer y Bofin (2008) desarrollaron un test de bondad de ajuste para evaluar el modelo *logit* multinomial. El test se basa en la estrategia de ordenar las observaciones de acuerdo al complemento de la probabilidad estimada de la categoría de referencia $(1-\hat{\pi}_{i0})$. Se forman g grupos, de aproximadamente n/g observaciones cada uno, y en cada grupo se calcula la suma de las frecuencias estimadas y observadas para cada categoría de respuesta,

$$O_{kj} = \sum_{l\in\Omega_k} \tilde{y}_{lj}$$

$$E_{kj} = \sum_{l\in\Omega_k} \hat{\pi}_{lj}$$

con $\tilde{y}_{lj}$ variable indicatriz, tal que $\tilde{y}_{lj}=1$ si $y_i=j$ y $\tilde{y}_{lj}=0$ en otro caso; $k=1,...,g$; $j=0,...,c-1$, c son las categorías de respuesta de Y; y Ω_k son las observaciones en el grupo k.

Las frecuencias estimadas y observadas pueden ser tabuladas en una tabla de contingencia utilizando los grupos como filas y las categorías como columnas, como se muestra en la Tabla 2.2.

Tabla 2.2: Tabla de contingencia de las frecuencias observadas (O_{kj}) y estimadas (E_{kj})

Grupo	Y=0	Y=1	$\cdots$	Y=c-1
1	O_{10} E_{10}	O_{11} E_{11}	$\cdots$	$O_{1,c-1}$ $E_{1,c-1}$
2	O_{10} E_{10}	O_{11} E_{11}	$\cdots$	$O_{1,c-1}$ $E_{1,c-1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
g	O_{g0} E_{g0}	O_{g1} E_{g1}	$\cdots$	$O_{g,c-1}$ $E_{g,c-1}$

Fuente: Fagerland *et al.* (2008).

De esta manera, el estadístico del test de bondad de ajuste multinomial queda definido como el estadístico chi-cuadrado de Pearson de la tabla de contingencia de $g\times c$:

$$C_g = \sum_{k=1}^{g}\sum_{j=0}^{c-1} \frac{(O_{kj}-E_{kj})^2}{E_{kj}}$$

Bajo la hipótesis nula de que el ajuste del modelo es el correcto y con un tamaño de muestra suficientemente grande, la distribución de C_g es $\chi^2_{(g-2)\times(c-1)}$.

Fagerland *et al.* (2008) mostraron mediante simulaciones que C_g tiene baja potencia para muestras de 100 observaciones, pero tiene potencia satisfactoria con 400 observaciones. Es decir, con 400 casos el test es capaz de detectar discrepancias entre un ajuste incorrecto y el modelo verdadero. Las simulaciones se llevaron a cabo utilizando un modelo *logit* multinomial con 3 categorías de respuesta, una covariable continua y $g = 8, 10, 12$. Los autores concluyen que la elección de la categoría de referencia no tiene o tiene muy poca incidencia en la distribución de C_g bajo la hipótesis nula. Sin embargo, para una base de datos en particular, C_g puede producir diferentes valores p según la categoría de referencia, pero esta diferencia no es muy grande.

El cálculo de esta prueba de bondad de ajuste fue desarrollado en Stata por Fagerland y Hosmer (2012).

2.2.4. Interpretación

La interpretación de los parámetros en la regresión logística multinomial es una extensión de la regresión logística binaria (Heeringa *et al.*, 2010; Hosmer *et al.*, 2013).

El *odds ratio* de $Y = j$ versus $Y = 0$ para el predictor p es

$$\widehat{OR}_{jp} = \exp(\hat{\beta}_{jp})$$

con $\hat{\beta}_{jp}$ parámetro estimado para la covariable p en el *logit* de j, y el intervalo de confianza es

$$IC(\widehat{OR}_{jp}) = \exp[\hat{\beta}_{jp} \pm t_{gl,1-\alpha/2} s(\hat{\beta}_{jp})] \quad .$$

Si en lugar de conocer el *odds ratio* de una categoría j con respecto a la categoría base, se quiere estimar el *odds ratio* con respecto a otra categoría j' se debe calcular

$$\widehat{OR}_{(jj')p} = \exp(\hat{\beta}_{jp} - \hat{\beta}_{j'p})$$

y

$$IC(\widehat{OR}_{(jj')p}) = \exp[(\hat{\beta}_{jp} - \hat{\beta}_{j'p}) \pm t_{gl,1-\alpha/2} s(\hat{\beta}_{jp} - \hat{\beta}_{j'p})] \quad .$$

Capítulo 3

Metodología

3.1. Datos

Los datos utilizados en esta investigación provienen de dos fuentes. Por un lado, se empleó la información correspondiente a los alumnos de la Facultad de Química del VII Censo de Estudiantes de Grado 2012 de la Universidad de la República, y por otro, se utilizaron los registros administrativos de la bedelía[1] de la misma Facultad.

Las dimensiones relevadas en el censo de estudiantes son: información sociodemográfica, vivienda, discapacidad, educación preuniversitaria, trabajo, carreras de grado, estudios terciarios no universitarios, estudios de posgrado, entornos virtuales de aprendizaje, lenguas, cogobierno, calidad de vida y actividades culturales.[2]

A partir de los registros administrativos se observó para cada estudiante censado en el 2012 la(s) carrera(s) en la(s) que se matriculó, el año de la última actividad académica y si egresó. Se entiende por actividad académica a la inscripción a curso o a examen, independientemente de su resultado. Con esta información se construyó la variable situación

[1] En Uruguay la bedelía corresponde al Departamento de Administración de la Enseñanza de cada Facultad y es la oficina que se encarga del registro y control de la actividad estudiantil; de la orientación general a estudiantes y docentes referente a trámites administrativos; de la inscripción de cursos, exámenes, trabajos experimentales y trabajos de extensión; de confeccionar las actas de exámenes y cursos; y de las tareas relacionadas con el egreso, entre otros cometidos.

[2] Para más información del censo de estudiantes de grado de la Udelar se puede consultar el documento realizado por la Dirección General de Planeamiento de la misma universidad: `http://planeamiento.udelar.edu.uy/files/2013/12/VII-Censo-de-Estudiantes-de-grado-2012.pdf`. Visitado el 30 de junio de 2016.

del estudiante al 31 de diciembre de 2015 (Y)[3], la cual se define como

$$Y = \{\text{egresó, abandonó, activo}\}$$

con:

Egresó. Si el estudiante egresó, entre el momento del censo y el 31 de diciembre de 2015, de al menos una de las carreras en las que se inscribió.

Abandonó. Si la última actividad académica del estudiante fue registrada en el año 2014 o antes, y esta actividad no fue con el fin de egresar.

Activo. Si el estudiante realizó al menos una actividad académica en el período comprendido entre el 1 de enero de 2015 y el 31 de diciembre del 2015, y esta actividad no fue con el fin de egresar.

3.2. Método

3.2.1. Modelo

En el modelo *logit* multinomial propuesto se consideró como variable respuesta la situación del estudiante definida en la sección anterior y como variables regresoras a aquellas que resultaron seleccionadas luego de realizar los pasos que se describen a continuación:

1. Se consultó a cuatro expertos en temas de rendimiento acádemico y abandono estudiantil de la Facultad de Química de la Udelar, tres docentes pertenecientes a la Unidad Académica de Educación Química y una docente perteneciente a la Secretaría de Apoyo al Estudiante. Cada uno de ellos señaló, entre todas las preguntas del censo, cuáles consideraba que era pertinente incorporar al modelo debido a que están relacionadas con la situación del estudiante.

2. Una vez obtenida la devolución de los expertos se tomaron en cuenta aquellas variables que fueron seleccionadas por los cuatro docentes consultados.

3. Se analizó mediante el test chi-cuadrado la posible asociación entre las variables del paso anterior y la situación del estudiante, teniendo en cuenta que la asociación fue significativa según el criterio de Hosmer y Lemeshow ($p < 0{,}25$).

4. Las variables introducidas en el modelo fueron las seleccionadas en el punto 3 y que no eran colineales entre ellas.

[3] De aquí en más se hará referencia a la variable Y como situación del estudiante, sin realizar la aclaración de que es al 31 de diciembre de 2015.

3.2.2. Población

La población objetivo de este trabajo estuvo formada por 4541 estudiantes que fueron censados en el VII Censo de Estudiantes de Grado 2012 y que se encontraban inscriptos, en el momento del censo, en al menos una carrera del Plan 2000 de la Facultad de Química de la Udelar.

Una vez definida la población, se establecieron los estratos y conglomerados a utilizar en el plan de muestreo. Para construir los estratos la población fue dividida en dos etapas. En la primera etapa, los estudiantes fueron separados en dos segmentos: uno que incluía a los estudiantes que se encontraban inscriptos sólo en carreras exclusivas de la Facultad de Química[4], y otro, que incluía a los estudiantes que se encontraban inscriptos en carreras de la Facultad de Química pero que son compartidas con otras facultades[5]. En la segunda etapa, el último segmento fue particionado en tres grupos: el primero, formado por los estudiantes que se encontraban inscriptos en Ingeniería Química, pero no en Ingeniería de Alimentos; el segundo, formado por los estudiantes inscriptos en Ingeniería de Alimentos, pero no en Ingeniería Química; y el tercero, formado por los estudiantes que se encontraban inscriptos tanto en Ingeniería Química como en Ingeniería de Alimentos.

Así, la población quedó dividida en cuatro estratos:

Estrato 1: incluye a los estudiantes que se encuentran inscriptos sólo en carreras exclusivas de la Facultad de Química (Químico Farmacéutico, Químico y Bioquímico Clínico).

Estrato 2: incluye a los estudiantes que se encuentran inscriptos en la carrera compartida con la Facultad de Ingeniería (Ingeniería Química), pero no son alumnos de la carrera compartida con las Facultades de Ingeniería, Veterinaria y Agronomía (Ingeniería de Alimentos).

Estrato 3: incluye a los estudiantes que se encuentran inscriptos en la carrera compartida con las Facultades de Ingeniería, Veterinaria y Agronomía (Ingeniería de Alimentos), pero no son alumnos de la carrera compartida con la Facultad de Ingeniería (Ingeniería Química).

Estrato 4: incluye a los estudiantes que se encuentran inscriptos tanto en la carrera compartida con la Facultad de Ingeniería (Ingeniería Química) como en la carrera compartida con las Facultad de Ingeniería, Veterinaria y Agronomía (Ingeniería de Alimentos).

[4]Químico Farmacéutico, Químico y Bioquímico Clínico

[5]Ingeniería Química, compartida con la Facultad de Ingeniería; e Ingeniería de Alimentos, compartida con las Facultades de Ingeniería, Veterinaria y Agronomía

En la Figura 3.1 se presenta un diagrama que ilustra la construcción de los estratos y en la Tabla 3.1 la distribución de los estudiantes según estrato. De acuerdo a la definición de estratos presentada en el párrafo anterior, en la Tabla 3.1 se observa que el 42 % de la población pertenece a las carreras exclusivas de Facultad de Química, el 32 % pertenece a Ingeniería Química, el 12 % a Ingeniería de Alimentos y el 14 % pertenece tanto a Ingeniería Química como a Ingeniería de Alimentos.

Figura 3.1: Construcción de estratos

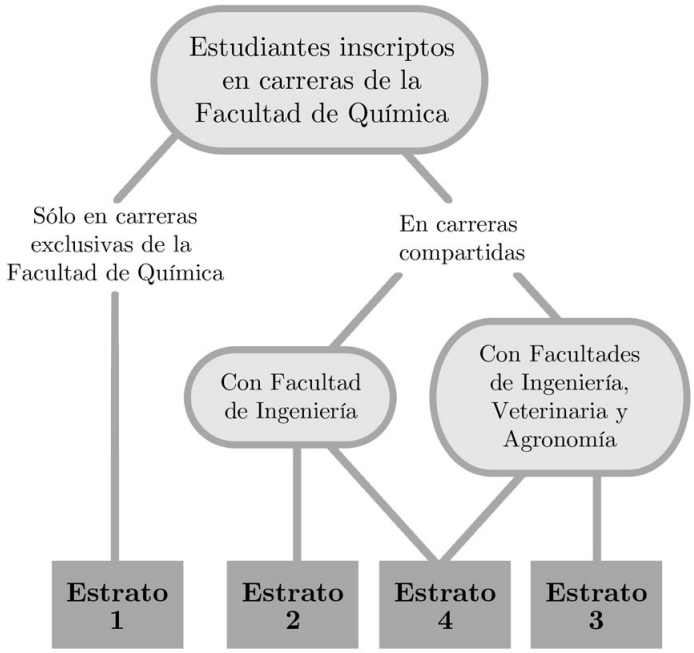


Tabla 3.1: Distribución de los estudiantes según estratos

Estrato	N	%
1	1904	41,9
2	1469	32,3
3	536	11,8
4	632	13,9
Total	4541	100

Con respecto a los conglomerados, estos fueron construidos a partir de dos carac-

terísticas de interés: el año de ingreso a la universidad y la cantidad de carreras en las que el estudiante se matriculó. Estas características fueron seleccionadas de acuerdo a la sugerencia de Levy y Lemeshow (2008) de que es posible reducir el error estándar de las estimaciones utilizando información auxiliar del conglomerado que puede estar correlacionada con la variable respuesta.

Debido a que en el año 2000 se crea un nuevo plan de estudios de las carreras en cuestión, a todos los estudiantes de los años anteriores se los considera como una generación, vale decir generación "Anterior al 2000". Asimismo existen 60 estudiantes que se encuentran inscritos en 4 o 5 carreras y se los agrupó con los estudiantes que se encuentran inscritos en 3. De esta manera, la población quedó dividida en 42 conglomerados, siendo el número 40 el más grande (361 estudiantes que están inscritos en una sola carrera y pertenecen a la generación 2012) y el número 42 el más pequeño (un estudiante que se encuentra inscrito en tres o más carreras y pertenece a la generación 2012) (Tabla 3.2).

Tabla 3.2: Distribución de los estudiantes según conglomerados: combinación de la generación y la cantidad de carreras con inscripción

Generación	Cantidad de carreras			Total
	1	2	3 o más	
Anterior al 2000	119 (1)	266 (2)	155 (3)	540
2000	50 (4)	35 (5)	20 (6)	105
2001	64 (7)	44 (8)	13 (9)	121
2002	89 (10)	49 (11)	32 (12)	170
2003	106 (13)	65 (14)	42 (15)	213
2004	140 (16)	109 (17)	38 (18)	287
2005	166 (19)	108 (20)	48 (21)	322
2006	229 (22)	128 (23)	32 (24)	389
2007	259 (25)	118 (26)	17 (27)	394
2008	235 (28)	123 (29)	19 (30)	377
2009	289 (31)	112 (32)	12 (33)	413
2010	317 (34)	111 (35)	16 (36)	444
2011	307 (37)	72 (38)	5 (39)	384
2012	361 (40)	20 (41)	1 (42)	382
Total	2731	1360	450	4541

Nota: entre paréntesis se indica el número de conglomerado

3.2.3. Muestreo

Una vez definidos los estratos y los conglomerados, se seleccionaron 1000 muestras para diez tamaños de muestra ($n = 100(100)1000$) bajo diferentes diseños de muestreo y en cada una de ellas, y en la población, se ajustaron modelos *logit* multinomiales de acuerdo a la definición del apartado 3.2.1. Es decir, la cantidad de muestras obtenidas para cada tipo de muestreo fue 10000. Los planes de muestreo utilizados fueron muestreo aleatorio simple, muestreo estratificado con asignación fija y asignación proporcional al tamaño, y muestreo por conglomerados en una etapa proporcional al tamaño.

En el muestreo estratificado con asignación fija se dividió el n en cuatro partes iguales, obteniendo en cada estrato la misma cantidad de observaciones.

Para el muestreo por conglomerados en cada extracción se seleccionaron cinco grupos con probabilidad de selección proporcional al tamaño y en cada grupo se eligieron cantidades iguales de observaciones ($n/5$).

El programa utilizado para el análisis de datos fue Stata 13, StataCorp (2013).

Capítulo 4

Resultados

Este capítulo se divide en dos secciones. En la primera, se muestran los resultados de cada uno de los pasos que se siguieron para seleccionar el modelo que luego se utilizó en los diferentes planes de muestreo; y en la segunda, se presenta el ajuste del modelo considerando diferentes tamaños de muestra y diferentes planes de muestreo.

En el Tabla 4.1 se presenta la distribución de la población en estudio según la situación del estudiante al 31 de diciembre de 2015 (N=4541). Se observa que aproximadamente dos tercios de los estudiantes se mantuvieron activos desde el momento de realización del censo (67 %), el 19 % abandonó los estudios y el 14 % logró egresar. Como se mencionó en el capítulo 3 esta será la variable respuesta a utilizar en el modelo *logit* multinomial y los datos para construirla se obtuvieron a partir del VII Censo de Estudiantes de Grado 2012 de la Udelar y de los registros administrativos de la Facultad de Química de la misma universidad.

Tabla 4.1: Distribución de la población según situación del estudiante

Situación del estudiante	%
Egresó	14,3
Abandonó	19,0
Activo	66,8
Total	100

4.1. Selección del modelo

A partir de la consulta realizada acerca de cuáles son las variables pertinentes para incorporar al modelo por la posible relación con la situación del estudiante, los expertos en rendimiento académico y abandono estudiantil de la Facultad de Química de la Udelar coincidieron en su selección en 8 de las 85 preguntas relevadas en el censo de estudiantes. Las ocho variables seleccionadas son las que se enumeran a continuación:

1. Sexo del estudiante
2. Región donde el estudiante cursó quinto año de enseñanza media
3. Tipo de institución donde el estudiante cursó quinto año de enseñanza media
4. Región donde el estudiante cursó sexto año de enseñanza media
5. Tipo de institución donde el estudiante cursó sexto año de enseñanza media
6. Si el estudiante tuvo que cambiar de lugar de residencia para poder desarrollar sus estudios universitarios
7. Horas que el estudiante trabaja en promedio por semana
8. Si por razones laborales el estudiante tuvo que realizar modificaciones en su trayectoria universitaria

Debido a que la región y el tipo de institución donde el estudiante cursó quinto año de enseñanza media y la región y el tipo de institución donde el estudiante cursó sexto año de enseñanza media son prácticamente iguales (97 % coincide la región donde cursó quinto y sexto, y 95 % coincide el tipo de institución), se optó por continuar trabajando con la región y el tipo de institución donde el estudiante cursó sexto año de enseñanza media. Se tomó esta decisión debido a que en Uruguay la aprobación del sexto año de enseñanza media es lo que habilita a una persona a ingresar a la educación superior, es decir, sexto es el último año de la enseñanza media.

De acuerdo al criterio de Hosmer y Lemeshow, al analizar individualmente la posible relación entre la situación del estudiante y las variables seleccionadas por los expertos, se encontró que ni el *sexo del estudiante* ($p = 0{,}75$) ni *si el estudiante tuvo que cambiar de lugar para realizar sus estudios universitarios* ($p = 0{,}40$) se encuentran asociados a ser un estudiante activo, que abandonó sus estudios o egresado. Sin embargo, las covariables *región* y *tipo de institución donde cursó sexto año de enseñanza media*, *horas de trabajo*

por semana y *por razones laborales tuvo que realizar modificaciones en su trayectoria universitaria* sí tienen una asociación significativa con la situación del estudiante ($p < 0{,}25$) (Tabla 4.2).

Tabla 4.2: Pruebas de asociación de cada una de las posibles variables regresoras con la situación del estudiante

	χ^2	p
Sexo	0,57	0,75
Si tuvo que cambiar de lugar de residencia para poder desarrollar sus estudios universitarios	1,85	0,40
Región donde cursó sexto año de educación media	43,88	0,04
Tipo de institución donde cursó sexto año de educación media	70,09	0,00
Horas que trabaja en promedio por semana	73,40	0,01
Si por razones laborales tuvo que realizar modificaciones en su trayectoria universitaria	36,78	0,00

Tomando los resultados obtenidos en las pruebas de asociación presentados en la Tabla 4.2, las variables regresoras a incluir en el modelo *logit* multinomial inicial son: región donde el estudiante cursó sexto año de enseñanza media (x_{reg}), tipo de institución donde el estudiante cursó sexto año de enseñanza media (x_{tipo}), horas que el estudiante trabaja en promedio semanalmente (x_{hs}) y si el estudiante tuvo que modificar su trayectoria universitaria por motivos laborales (x_{tray}). En las Tablas 4.3-4.6 se presenta la distribución de los estudiantes según la situación para cada una de estas covariables.

De acuerdo a los resultados, no parece haber diferencias claras en la situación del estudiante entre aquellos que provienen de una institución de Montevideo o una del Interior, aunque el porcentaje de egresados es levemente mayor para los que provienen de Montevideo que del Interior (15 % y 13 %, respectivamente) (Tabla 4.3).

Tabla 4.3: Población por región donde cursó educación media según situación del estudiante

Situación	Región		Total
	Montevideo	Interior	
Egresó	15,4	13,0	14,3
Abandonó	18,2	19,9	19,0
Activo	66,4	67,1	66,8
Total	100 (2480)	100 (2061)	100 (4541)

Nota: entre paréntesis se indica la cantidad de estudiantes

En cambio, la situación del estudiante es distinta según el tipo de institución de la que proviene el alumno. En la Tabla 4.4 se observa que el porcentaje de egresados es mayor entre aquellas personas que provienen de una institución privada (20 %) que entre aquellas que provienen del sistema público (11 %). Por el contrario, en la situación de abandono los porcentajes son al revés: 21 % de los estudiantes procedentes de la educación pública y 16 % de los estudiantes de la educación privada. La proporción de alumnos activos es 68 % y 64 % para las categorías pública y privada, respectivamente.

Tabla 4.4: Población por tipo de institución donde cursó educación media según situación del estudiante

Situación	Tipo de institución educación media		Total
	Pública	Privada	
Egresó	11,4	20,2	14,3
Abandonó	20,6	15,5	19,0
Activo	68,0	64,3	66,8
Total	100 (3064)	100 (1477)	100 (4541)

Nota: entre paréntesis se indica la cantidad de estudiantes

En lo que respecta a las horas semanales que trabajaba el estudiante al momento del censo, se destaca que el 59 % de los que trabajaban 40 horas o más se encontraba activo al 31 de diciembre de 2015, mientras que entre los que no trabajaban este porcentaje asciende a 72 %. El porcentaje de abandono es similar entre los que no trabajaban y los que trabajaban 20 horas o menos (17-18 %), y similar entre los que trabajaban de 21 a 40 horas y los que trabajaban más de 40 horas (20-21 %). A su vez, la cantidad relativa de egresados entre los trabajadores de más de 40 horas es el doble (20 %) que la cantidad relativa de egresados entre los estudiantes que no trabajaban al momento del censo (10 %).

La situación de los estudiantes que trabajaban 20 horas o menos y entre 21 y 40 horas es similar al del total de la población (Tabla 4.5).

Tabla 4.5: Población por horas que trabaja por semana según situación del estudiante

Situación	Horas que trabaja por semana				Total
	No trabaja	20 hs. o menos	Entre 21 y 40 hs.	Más de 40 hs.	
Egresó	10,0	16,1	15,2	20,3	14,3
Abandonó	17,6	17,0	20,3	20,7	19,0
Activo	72,3	67,0	64,5	59,1	66,8
Total	100 (1843)	100 (454)	100 (1252)	100 (992)	100 (4541)

Nota: entre paréntesis se indica la cantidad de estudiantes

En la Tabla 4.6 se observa que el porcentaje de alumnos que alcanzaron el egreso es menor entre aquellos que tuvieron que modificar su trayectoria universitaria por razones laborales (13 %) que entre aquellos que no la modificaron (21 %). Sin embargo, el porcentaje de abandono es similiar en ambas categorías, 19 % y 20 %. La proporción de estudiantes activos es 67 % para los que modificaron su trayectoria por motivos laborales y 61 % para los que no lo hicieron.

Tabla 4.6: Población por modificación de trayectoria universitaria por razones laborales según situación del estudiante

Situación	Modificación de trayectoria universitaria por razones laborales		Total
	No	Sí	
Egresó	20,5	12,8	15,6
Abandonó	18,8	20,4	19,8
Activo	60,7	66,9	64,6
Total	100 (1280)	100 (2212)	100 (3492)[a]

Notas: Entre paréntesis se indica la cantidad de estudiantes.
[a] Se excluyen 1049 estudiantes que nunca trabajaron al momento del censo.

Tomando las covariables seleccionadas y la variable de respuesta Y, presentada en el capítulo metodológico 3, el modelo se especifica considerando *abandono* como categoría de referencia. De esta manera, el modelo *logit* multinomial es el siguiente:

$$\begin{aligned} g_{Egr}(\mathbf{x}) &= ln\left[\frac{P(Y = Egreso|\mathbf{x})}{P(Y = Abandono|\mathbf{x})}\right] \\ &= \beta_{10} + \beta_{1reg}x_{reg} + \beta_{1tipo}x_{tipo} + \beta_{1hs}x_{hs} + \beta_{1tray}x_{tray} \\ &= \mathbf{x}^\top\beta_{\mathbf{1}} \end{aligned} \tag{4.1}$$

$$\begin{aligned} g_{Act}(\mathbf{x}) &= ln\left[\frac{P(Y = Activo|\mathbf{x})}{P(Y = Abandono|\mathbf{x})}\right] \\ &= \beta_{20} + \beta_{2reg}x_{reg} + \beta_{2tipo}x_{tipo} + \beta_{2hs}x_{hs} + \beta_{2tray}x_{tray} \\ &= \mathbf{x}^\top\beta_{\mathbf{2}} \end{aligned} \tag{4.2}$$

con

x_{reg}: Región donde el estudiante cursó el último año de enseñanza media (Montevideo / Interior),

x_{tipo}: Tipo de institución donde el estudiante cursó el último año de enseñanza media (Pública / Privada),

x_{hs}: Horas que el estudiante trabaja en promedio semanalmente (No trabaja / 20 hs. o menos / Entre 21 y 40 hs. / Más de 40 hs.) y

x_{tray}: Si el estudiante tuvo que modificar su trayectoria universitaria por motivos laborales (Sí / No).

Es decir, todas las covariables son de escala nominal, tres son binarias (x_{reg}, x_{tipo} y x_{tray}) y una tiene cuatro categorías (x_{hs}).

Así, las probabilidades condicionales de cada situación del estudiante (*Abandonó*, *Egresó*, *Activo*) dadas las covariables son

$$P(Y = Abandono|\mathbf{x}) = \frac{1}{1 + e^{g_{Egr}(\mathbf{x})} + e^{g_{Act}(\mathbf{x})}} \quad ,$$

$$P(Y = Egreso|\mathbf{x}) = \frac{e^{g_{Egr}(\mathbf{x})}}{1 + e^{g_{Egr}(\mathbf{x})} + e^{g_{Act}(\mathbf{x})}} \quad \text{y}$$

$$P(Y = Activo|\mathbf{x}) = \frac{e^{g_{Act}(\mathbf{x})}}{1 + e^{g_{Egr}(\mathbf{x})} + e^{g_{Act}(\mathbf{x})}} \quad .$$

Una vez especificado el modelo, las estimaciones con los datos poblacionales se obtienen con 4 iteraciones.

En la Tabla 4.7 se presentan los coeficientes estimados, su error estándar, el estadístico de Wald y la probabilidad de significación p. Allí se puede observar que la región no resulta ser una variable significativa ni en el *logit egresó* ($p = 0{,}09$) ni en el *logit activo* para $\alpha = 0{,}05$ ($p = 0{,}36$), lo cual tiene concordancia con el análisis bivariado de la Tabla 4.3. El resto de los coeficientes asociados a la covariables fueron significativos en *egresó* pero no en *activo*, con excepción de la categoría trabaja más de 40 hs. y la constante que resultaron significativas en ambos *logits*. A partir de estos resultados, se decidió estimar un nuevo modelo sin considerar la variable región.

Tabla 4.7: Estimación de los coeficientes del modelo con todas las posibles covariables

Logit / Variable[a]	$\hat{\beta}$	$s(\hat{\beta})$	z	p
Egresó				
reg(int)	0,225	0,134	1,68	0,09
tipo(priv)	0,954	0,141	6,77	$< 0{,}001$
hs(<20)	0,761	0,218	3,49	$< 0{,}001$
hs(20-40)	0,694	0,177	3,93	$< 0{,}001$
hs(> 40)	1,022	0,183	5,58	$< 0{,}001$
tray(sí)	-0,715	0,128	-5,58	$< 0{,}001$
constante	-0,937	0,178	-5,25	$< 0{,}001$
Activo				
reg(int)	-0,089	0,096	-0,92	0,36
tipo(priv)	0,105	0,109	0,96	0,34
hs(<20)	0,068	0,158	0,43	0,67
hs(20-40)	-0,160	0,122	-1,32	0,19
hs(> 40)	-0,277	0,131	-2,12	$< 0{,}05$
tray(sí)	0,104	0,099	1,05	0,29
constante	1,252	0,120	10,41	$< 0{,}001$

N=3492, $\chi^2_{12} = 180{,}54$, $p < 0{,}001$

[a] Las categorías de referencia para las covariables son: reg(Mvd), tipo(púb), hs(no trabaja) y tray(no)

Las covariables incluidas en el nuevo modelo (x_{tipo}, x_{hs} y x_{tray}) resultaron significativas para el *logit egresó* pero no para *activo*, con excepción de la categoría trabaja más de 40 hs. que es significativa en ambos *logits* (Tabla 4.8). Por esta razón, se estimaron nuevos modelos quitando las covariables de a una y se analizó la variación de $\hat{\beta}$. De acuerdo al criterio de Hosmer *et al.* (2013) no es recomendable excluir una variable del modelo si esto

causa una variación de más del 20-25 % en las estimaciones de los coeficientes del resto de las covariables.

Tabla 4.8: Estimación de los coeficientes del modelo sin considerar la variable región

Logit / Variable	$\hat{\beta}$	$s(\hat{\beta})$	z	p
Egresó				
tipo(priv)	0,842	0,124	6,79	$< 0{,}001$
hs(<20)	0,741	0,218	3,41	$< 0{,}01$
hs(20-40)	0,674	0,176	3,83	$< 0{,}001$
hs(> 40)	1,010	0,183	5,53	$< 0{,}001$
tray(sí)	-0,723	0,128	5,66	$< 0{,}001$
constante	-0,780	0,151	5,16	$< 0{,}001$
Activo				
tipo(priv)	0,147	0,099	1,48	0,14
hs(<20)	0,073	0,158	0,46	0,65
hs(20-40)	-0,154	0,122	-1,27	0,21
hs(> 40)	-0,273	0,131	-2,09	$< 0{,}05$
tray(sí)	0,110	0,099	1,11	0,27
constante	1,190	0,101	11,76	$< 0{,}001$

N=3492, $\chi^2_{12} = 172{,}92$, $p < 0{,}001$

[a] Las categorías de referencia para las covariables son: tipo(púb), hs(no trabaja) y tray(no)

En la Tabla 4.9 se puede observar que al eliminar del modelo el *tipo de institución de enseñanza media*, los coeficientes asociados a las otras variables regresoras varían menos de 11 % en los dos *logits*, con excepción de la constante. Si se extrae del modelo la variable *horas que el estudiante trabaja*, los coeficientes asociados a la modificación de trayectoria universitaria varían 32 % y 76 % en *egresó* y *activo*, respectivamente. Al estimar el modelo sin la variable que indica *si el estudiante tuvo que modificar su trayectoria universitaria por motivos laborales*, los coeficientes asociados al tipo de institución y a las horas que trabaja por semana varían entre 27 % y 144 %.

Tabla 4.9: Comparación de los coeficientes estimados entre el modelo completo y los reducidos

	$\hat{\beta}$	$\hat{\beta}_{-x_{tipo}}$	$\Delta\beta\,\%_{-x_{tipo}}$	$\hat{\beta}_{-x_{hs}}$	$\Delta\beta\,\%_{-x_{hs}}$	$\hat{\beta}_{-x_{tray}}$	$\Delta\beta\,\%_{-x_{tray}}$
Egresó							
tipo(int)	0,842			0,860	2.1	0,887	5,3
hs(<20)	0,741	0,774	4,4			0,540	-27,2
hs(20-40)	0,674	0,676	0,3			0,349	-48,2
hs(> 40)	1,010	1,040	2,9			0,605	-40,1
tray(sí)	-0,723	-0,799	10,5	-0,490	-32,2		
constante	-0,780	-0,456	-41,5	-0,255	-67,3	-0,924	18,5
Activo							
tipo(int)	0,147			0,149	1,6	0,208	41,7
hs(<20)	0,073	0,077	5,8			-0,032	-143,7
hs(20-40)	-0,154	-0,155	0,7			-0,238	54,6
hs(> 40)	-0,273	-0,271	-0,07			-0,348	27,3
tray(sí)	0,110	0,100	-9,3	0,027	-75,7		
constante	1,190	1,239	3,8	1,124	-5,8	1,343	12,6

No corresponde

De acuerdo a los resultados anteriores y siguiendo la recomendación de Hosmer *et al.* (2013) el modelo seleccionado es el que incluye la *cantidad de horas que trabaja el estudiante* y *si tuvo que modificar la trayectoria universitaria por motivos laborales*, y no considera la *región* ni el *tipo de institución donde el estudiante cursó el último año de enseñanza media.*

Una vez especificado el modelo con dos covariables (*cantidad de horas que trabaja el estudiante* y *si tuvo que modificar la trayectoria universitaria por motivos laborales*), se analizó la posible interacción entre ellas. Los resultados mostraron que de las cuatro interacciones de cada *logit*, dos son significativas en *Egresó* (hs(20-40)*tray(sí) y hs(> 40)*tray(sí)) y ninguna en *Activo.*

A pesar de lo anterior, se decide seleccionar el modelo sin interacciones por dos razones. Por un lado, en este trabajo se estimará el modelo en miles de muestras con diferentes diseños de muestreo, y el costo computacional de incorporar los términos de interacción es de gran magnitud; y por otro, no es el objetivo de la investigación proponer un modelo adecuado para analizar la situación del estudiante, sino que es analizar las estimaciones bajo diferentes planes de muestreo.

En las ecuaciones (4.3) y (4.4) se presenta la especificación del modelo seleccionado. El test de razón de verosimilitud de bondad de ajuste dio como resultado que este puede

ser un modelo adecuado ($\chi^2_8 = 115{,}9$, $p < 0{,}001$).

$$\begin{aligned} g_{Egr}(\mathbf{x}) &= ln\left[\frac{P(Y = egreso|\mathbf{x})}{P(Y = abandono|\mathbf{x})}\right] \\ &= -0{,}46 + 0{,}77x_{hs(<20)} + 0{,}068x_{hs(20-40)} + 1{,}04x_{hs(>40)} - 0{,}80x_{tray} \end{aligned} \tag{4.3}$$

$$\begin{aligned} g_{Act}(\mathbf{x}) &= ln\left[\frac{P(Y = activo|\mathbf{x})}{P(Y = abandono|\mathbf{x})}\right] \\ &= 1{,}24 - 0{,}07x_{hs(<20)} - 0{,}16x_{hs(20-40)} - 0{,}27x_{hs(>40)} - 0{,}10x_{tray} \end{aligned} \tag{4.4}$$

Los *odds ratio* para el modelo seleccionado se presentan en la Tabla 4.10. El *odds* de haber egresado versus haber abandonado es aproximadamente dos entre los alumnos que trabajaban 40 horas o menos con respecto a los que no trabajaban ($\hat{OR}_{Egr:hs(<20)} = 2{,}2$; $\hat{OR}_{Egr:hs(20-40)} = 2{,}0$), mientras que, es casi tres para aquellos que trabajaban más de 40 horas ($\hat{OR}_{Egr:hs(40)} = 2{,}8$). Por otra parte, el *odds* de haber egresado versus haber abandonado es significativamente bajo para los alumnos que tuvieron que modificar su trayectoria universitaria en relación a los que no lo hicieron ($\hat{OR}_{Egr:tray} = 0{,}45$). En relación a los estudiantes que no trabajaban, el *odds* de mantenerse activo o abandonar es 0,8 para los alumnos que estaban ocupados más de 40 horas.

Estos resultados sugieren que el trabajar puede ser un factor que incentive a los alumnos a continuar con sus estudios y alcanzar el egreso. Sin embargo, para arribar a esta conclusión se deberían realizar estudios enfocados en la temática de egreso y situación laboral, debido a que no es el cometido de este trabajo.

Tabla 4.10: Estimación de los *odds ratio* del modelo seleccionado

Logit / Variable	$\hat{OR}$	$s(\hat{OR})$	z	p	IC del 95 %
Egresó					
hs(<20)	2,17	0,469	3,58	< 0,001	1,42; 3,31
hs(20-40)	1,97	0,344	3,86	< 0,001	1,40; 2,77
hs(> 40)	2,83	0,514	5,72	< 0,001	1,98; 4,04
tray(sí)	0,45	0,057	-6,31	< 0,001	0,35; 0,58
constante	0,63	0,090	-3,21	< 0,01	0,48; 0,84
Activo					
hs(<20)	1,08	0,171	0,49	0,63	0,79; 1,47
hs(20-40)	0,86	0,104	-1,27	0,20	0,67; 1,09
hs(> 40)	0,76	0,100	-2,08	< 0,05	0,59; 0,98
tray(sí)	1,10	0,109	1,01	0,31	0,91; 1,34
constante	3,45	0,335	12,77	< 0,001	2,85; 4,17

N=3492, $\chi^2_8 = 115{,}9$, $p < 0{,}001$

[a] Las categorías de referencia para las covariables son: hs(no trabaja) y tray(no)

Por último, se evaluó el modelo seleccionado a través del test de bondad de ajuste propuesto por Fagerland *et al.* (2008) y presentado en el apartado 2.2.3 del capítulo 2. La cantidad de grupos, g, por defecto que utiliza la prueba desarrollada por Fagerland y Hosmer (2012) en Stata es 10. Para este modelo, con los datos poblacionales, se logró calcular empleando seis grupos o menos, $g \leq 6$. De lo contrario, algunas de las celdas de la tabla de contingencia considerada para calcular el estadístico quedaban sin observaciones.

A partir de los resultados del test de Fagerland, para un nivel de significación de 0,05, se considera que el modelo es adecuado con $g = 3$ ($\chi^2_2 = 3{,}72$, $p = 0{,}16$) y $g = 5$ ($\chi^2_6 = 12{,}04$, $p = 0{,}06$). Sin embargo, se concluye lo contrario si la prueba es realizada con $g = 4$ ($\chi^2_4 = 11{,}49$, $p = 0{,}02$) y $g = 6$ ($\chi^2_8 = 23{,}81$, $p = 0{,}002$).

En resumen, si se considera el test de razón de verosimilitud, el modelo resulta adecuado. No obstante, para llegar a la misma conclusión utilizando la prueba de Fagerland depende de la cantidad de grupos que se empleen.

Dado que el objetivo de este trabajo no es proponer un modelo adecuado para analizar la situación del estudiante, sino que es analizar las estimaciones del modelo *logit* multinomial bajo diferentes diseños de muestreo, se utilizará el modelo seleccionado en lo que sigue a continuación.

4.2. Estimación del modelo bajo diferentes planes de muestreo

Esta sección se divide en tres apartados y en cada uno de ellos se presentan los resultados para los diferentes tamaños de muestras y tipos de muestreo. En el primer apartado se expone la cantidad de observaciones que efectivamente se utilizaron para estimar los modelos, en el segundo se muestran las estimaciones de uno de los coeficientes y de los errores estándares del coeficiente, y en el tercero se presenta el desempeño del estadístico de bondad de ajuste C_g.

4.2.1. Tamaño de muestra efectivo

La pérdida de observaciones del modelo se debe a que la pregunta sobre modificación de trayectoria universitaria por motivos labores excluye a los estudiantes que nunca habían trabajado al momento del censo. Por esta razón, la cantidad de unidades empleadas para estimar el modelo en cada una de las muestras extraídas siempre fue menor al tamaño propuesto.

Se llamó n propuesto al tamaño de muestra propuesto para seleccionar en las extracciones ($n = 100(100)1000$) y n efectivo a la cantidad de observaciones que efectivamente se utilizó para estimar el modelo una vez obtenida la muestra.

En la Tabla 4.11 se presentan los estadísticos descriptivos del n efectivo según el tipo de muestreo y n propuesto. Los resultados mostraron que bajo MAS y muestreo estratificado con asignación fija y asignación proporcional al tamaño, en promedio, la pérdida de observaciones es aproximadamente 23 % para todos los tamaños de muestra requeridos, lo cual coincide con la pérdida de observaciones que se utilizaron para estimar el modelo en la población. En cambio, con muestreo por conglomerados la pérdida de observaciones, en promedio, varía entre 24 % (n propuesto menor o igual a 600) y 38 % (n propuesto=1000).

Si se analiza el coeficiente de variación, el comportamiento es igual en los cuatro tipo de muestreo considerados: a medida que el tamaño de muestra propuesto aumenta, la variabilidad es menor, es decir, aumenta la homogeneidad. El coeficiente de variación del n efectivo es similar entre el MAS y ambos tipos de muestreo estratificado, el cual toma valores desde 0,014 hasta 0,056. Mientras que este coeficiente se sitúa entre 0,009 y 0,045 bajo muestreo por conglomerados.

Tabla 4.11: Tamaño de muestra efectivo para estimar el modelo por tipo de muestreo y tamaño de muestra propuesto

Muestreo	n propuesto	n efectivo				
		Media	Error	CV	Mín	Máx
MAS	100	77	3.996	0,052	65	90
	200	153	5.949	0,039	134	171
	300	230	7.107	0,031	207	251
	400	307	8.139	0,027	282	333
	500	383	9.127	0,024	353	411
	600	461	10.030	0,022	425	488
	700	538	10.721	0,020	494	566
	800	615	11.194	0,018	574	645
	900	691	11.747	0,017	651	722
	1000	768	12.026	0,016	731	800
Estratificado con asignación fija	100	77	4.010	0,052	64	90
	200	154	5.571	0,036	138	172
	300	231	6.871	0,030	211	252
	400	308	7.937	0,026	287	334
	500	384	8.487	0,022	354	412
	600	462	9.606	0,021	430	491
	700	538	10.317	0,019	501	573
	800	616	10.332	0,017	587	647
	900	692	10.492	0,015	657	724
	1000	769	11.081	0,014	733	805
Estratificado con asignación proporcional	100	77	4.320	0,056	62	88
	200	154	5.664	0,037	137	170
	300	231	7.012	0,030	204	252
	400	308	8.054	0,026	280	331
	500	385	8.985	0,023	356	409
	600	461	9.837	0,021	429	489
	700	538	10.384	0,019	506	573
	800	615	10.799	0,018	582	648
	900	692	11.389	0,016	651	727
	1000	769	11.825	0,015	724	805
Por conglomerados	100	76	3.340	0,045	65	86
	200	153	4.432	0,029	138	169
	300	229	5.326	0,023	213	248
	400	306	5.718	0,019	286	324
	500	381	6.126	0,016	362	402
	600	458	6.339	0,014	440	479
	700	524	6.061	0,012	505	544
	800	577	6.116	0,011	559	595
	900	602	5.883	0,010	584	622
	1000	622	5.740	0,009	604	640

Notas: (1) n propuesto corresponde al tamaño de la muestra seleccionada y n efectivo a la cantidad de observaciones que efectivamente se utilizó para estimar el modelo.

(2) Error: error estándar, CV: coeficiente de variación, Mín: mínimo y Máx: máximo.

(3) La pérdida de observaciones del modelo se debe a que la pregunta sobre modificación de trayectoria universitaria por motivos labores excluye a los estudiantes que nunca trabajaron al momento del censo.

4.2.2. Estimación de los coeficientes y de los errores estándares

Dado que al trabajar con toda la población, la convergencia en las estimaciones del modelo se logró con cuatro iteraciones, en los diferentes planes de muestreo se indicó al programa realizar hasta 15 iteraciones. En casi todos los casos se logró la convergencia, con excepción de dos muestras de tamaño 100 en el muestreo por conglomerados.

Antes de presentar los resultados de las estimaciones se debe realizar la siguiente aclaración. Como se vio en la sección 4.1, el ajuste del modelo seleccionado implica la estimación de diez coeficientes (cuatro β's más la constante por cada *logit*). Los resultados obtenidos para cada uno de los $\hat{\beta}$'s son los mismos, por esta razón y con el fin de simplificar la lectura, se mostrará el comportamiento de un solo $\hat{\beta}$.

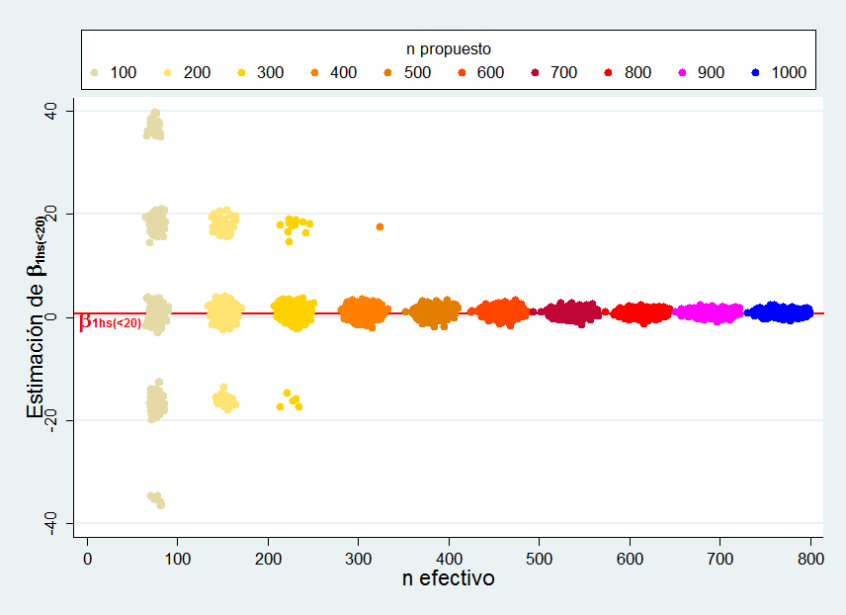
En las Figuras 4.1, 4.3, 4.5 y 4.7 se presentan las estimaciones en el *logit egresó* del coeficiente asociado a la categoría trabaja 20 horas o menos ($\hat{\beta}_{1hs(<20)}$). Los gráficos muestran $\hat{\beta}_{1hs(<20)}$ para cada uno de los tipos de muestreo utilizados y para las 1000 muestras obtenidas para cada n ($n = 100(100)1000$), según el tamaño de muestra propuesto y el tamaño de muestra efectivo, y el valor poblacional del coeficiente ($\beta_{1hs(<20)} = 0{,}77$).

En las estimaciones de las Figuras 4.1, 4.5 y 4.7 se observan valores de $\hat{\beta}_{1hs(<20)}$ extremos para $n \leq 400$, con MAS, muestreo estratificado con asignación proporcional y muestreo por conglomerados. Mientras que en el muestreo estratificado con asignación fija los valores extremos de $\hat{\beta}_{1hs(<20)}$ se producen para $n \leq 300$ (Figura 4.3). En las muestras con esta característica se da tanto una sobreestimación de los $\beta_{1hs(<20)}$, con valores por encima de 10, así como también una subestimación de los $\beta_{1hs(<20)}$, con valores por debajo de -10.

Al analizar los casos donde se producen $\hat{\beta}_{1hs(<20)}$ extremos, se encontró que la sobreestimación se genera en aquellas muestras donde ninguna o solo una observación cumple con la condición de ser un estudiante que abandonó sus estudios y que trabaja 20 hs. o menos por semana. Del mismo modo, la subestimación se produce a partir de muestras donde una o ninguna observación cumple con la condición de ser un estudiante que egresó y que trabaja 20 hs. o menos por semana.

Las estimaciones parecen ser consistentes a partir de n propuesto ≥ 800 en el caso de MAS, n propuesto ≥ 900 en el caso del muestreo estratificado y n propuesto=1000 en el muestreo por conglomerados.

Figura 4.1: $\hat{\beta}_{1hs(<20)}$ bajo MAS para diferentes tamaños de muestra



Con respecto a la distribución de las estimaciones, de acuerdo a la teoría de los modelos *logit* multinomial, $\hat{\beta}_{1hs(<20)}$ tiene una distribución normal. Por lo tanto, para evaluar si esto se cumple se realizaron los gráficos de las estimaciones y se calculó el test de normalidad de Shapiro Wilk para cada n propuesto y tipo de muestreo.

Al observar la distribución de $\hat{\beta}_{1hs(<20)}$ bajo MAS se podría decir que para los $n \geq 400$ las curvas se asemejan a una distribución normal (Figura 4.2). Sin embargo, mediante el test de normalidad de Shapiro Wilk no se rechaza la distribución normal en muestras con n propuesto mayor o igual a 800, para un nivel de significación de 0,05 ($n = 100(100)700$ ($p <$ 0,05), $n = 800$ ($p = 0{,}11$), $n = 900$ ($p = 0{,}94$) y $n = 900$ ($p = 0{,}84$)).

De manera similar a lo que ocurre bajo MAS, el análisis gráfico de la distribución de $\hat{\beta}_{1hs(<20)}$ para muestreo estratificado con asignación fija muestra que la distribución normal parece ocurrir a partir de $n \geq 400$ (Figura 4.4). Asimismo, en este tipo de muestreo y considerando $\alpha = 0{,}05$, mediante la prueba de hipótesis no se rechaza la distribución normal de $\hat{\beta}_{1hs(<20)}$ para $n \geq 500$, con excepción de $n = 700$ ($p = 0{,}01$): $n = 500$ ($p = 0{,}82$), $n = 600$ ($p = 0{,}10$), $n = 800$ ($p = 0{,}17$), $n = 900$ ($p = 0{,}19$) y $n = 1000$ ($p = 0{,}84$).

En la Figura 4.6 se observa que bajo muestreo estratificado con asignación proporcio-

Figura 4.2: Distribución de $\hat{\beta}_{1hs(<20)}$ bajo MAS para diferentes tamaños de muestra

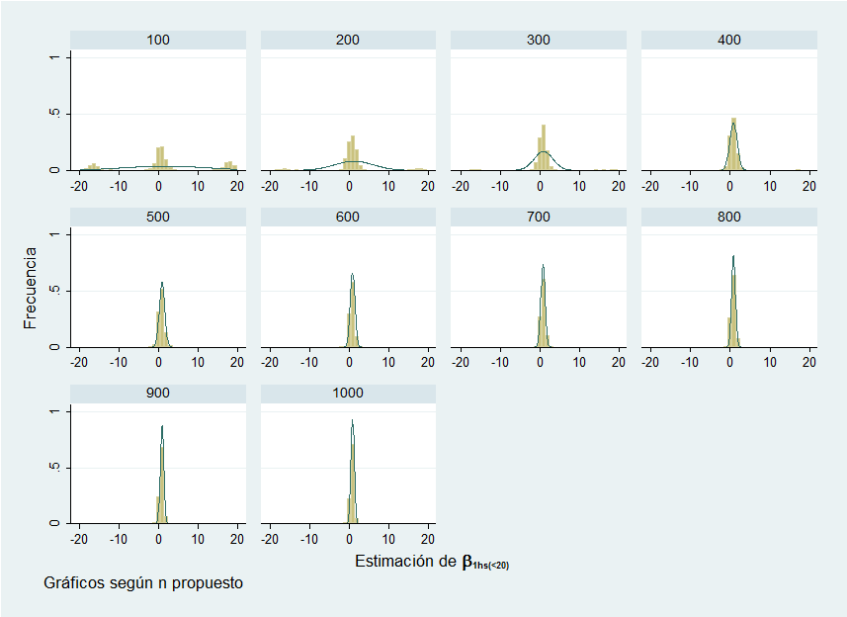


Figura 4.3: $\hat{\beta}_{1hs(<20)}$ bajo muestreo estratificado con asignación fija para diferentes tamaños de muestra

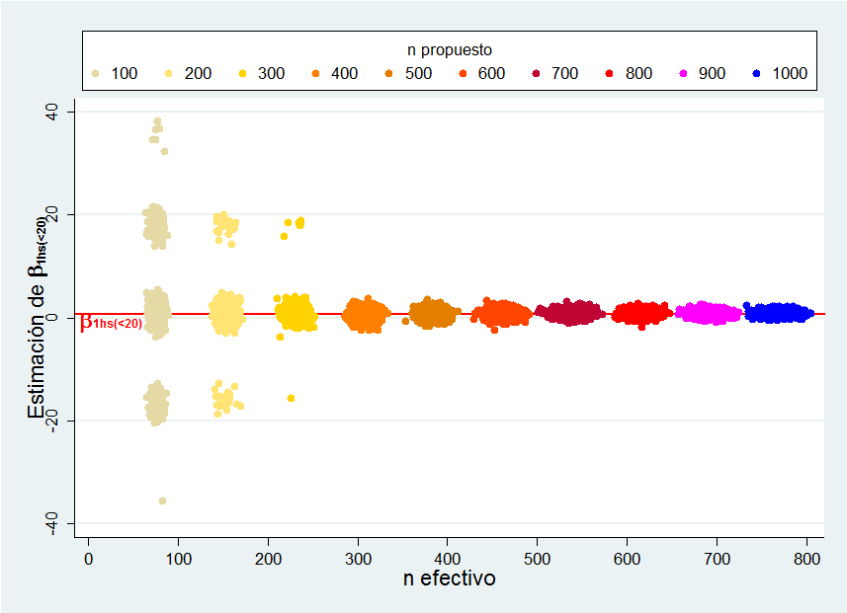
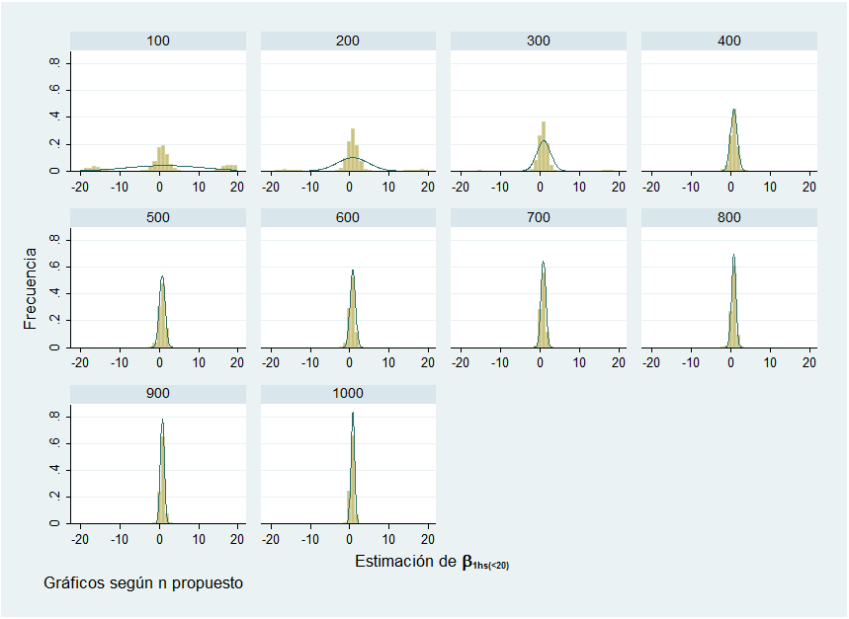


Figura 4.4: Distribución de $\hat{\beta}_{1hs(<20)}$ bajo muestreo estratificado con asignación fija para diferentes tamaños de muestra



nal al tamaño, la distribución de $\hat{\beta}_{1hs(<20)}$ parece asimilarse a una Normal para tamaños de muestra iguales o mayores a 500. No obstante, al obtener los resultados de las pruebas de normalidad con $\alpha = 0{,}05$ no se rechaza la distribución normal de $\hat{\beta}_{1hs(<20)}$ para $n = 800$ ($p = 0{,}18$), $n = 900$ ($p = 0{,}71$) y $n = 1000$ ($p = 0{,}90$).

A diferencia de lo que ocurre bajo MAS y muestreo estratificado, la distribución de $\hat{\beta}_{1hs(<20)}$ en el muestreo por conglomerados no se visualiza gráficamente de forma clara (Figura 4.8). Esto es concordante con los test de Shapiro Wilk realizados con este muestreo. Considerando un nivel del significación de 5 % se rechaza la distribución normal para la mayoría de los tamaños de muestra, con excepción de $n = 600$ ($p = 0{,}12$), $n = 900$ ($p = 0{,}62$) y $n = 1000$ ($p = 0{,}35$).

En resumen, de acuerdo a los tests de hipótesis realizados con $\alpha = 0{,}05$, no se rechaza la distribución Normal de $\hat{\beta}_{1hs(<20)}$ para muestras con n propuesto de al menos 800 observaciones bajo MAS y muestreo estratificado, y para muestras con al menos 900 observaciones bajo muestreo por conglomerados.

Figura 4.5: $\hat{\beta}_{1hs(<20)}$ bajo muestreo estratificado con asignación proporcional al tamaño para diferentes tamaños de muestra

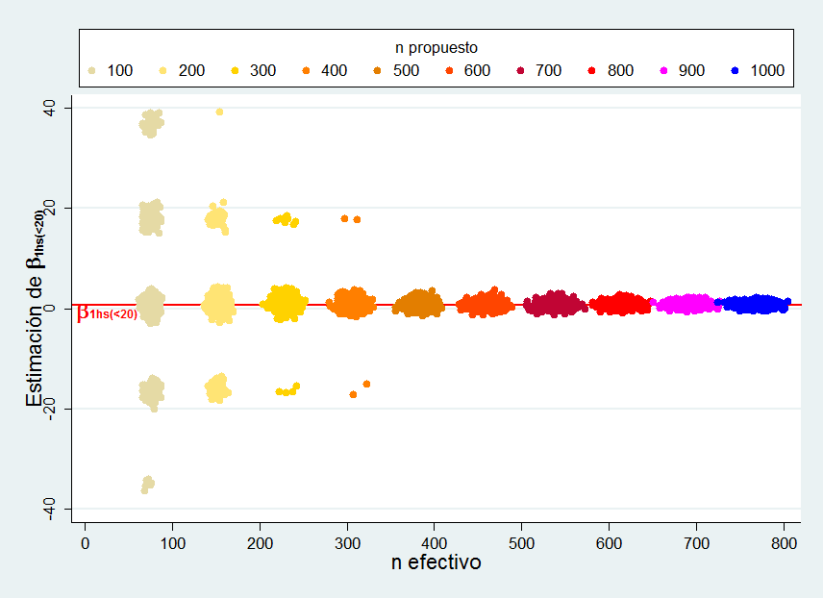


Figura 4.6: Distribución de $\hat{\beta}_{1hs(<20)}$ bajo muestreo estratificado con asignación proporcional al tamaño para diferentes tamaños de muestra

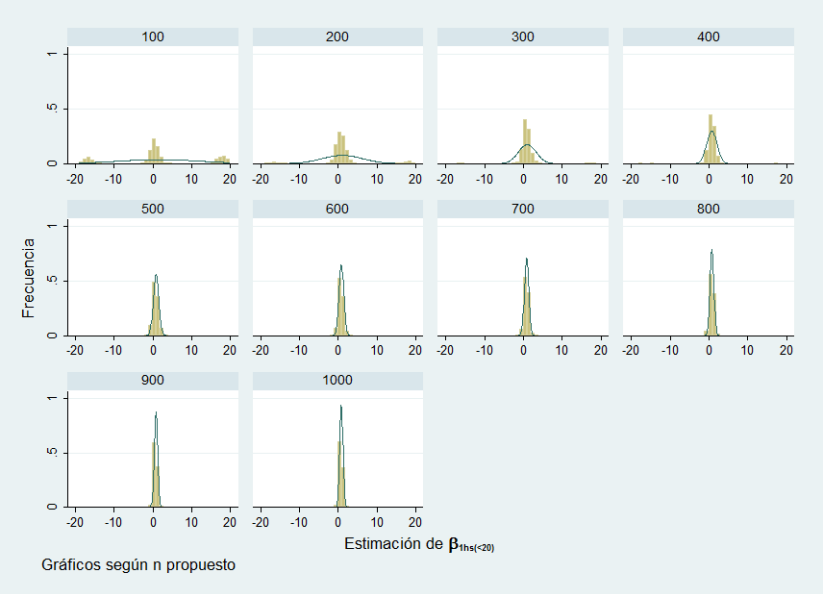


Figura 4.7: $\hat{\beta}_{1hs(<20)}$ bajo muestreo por conglomerados para diferentes tamaños de muestra

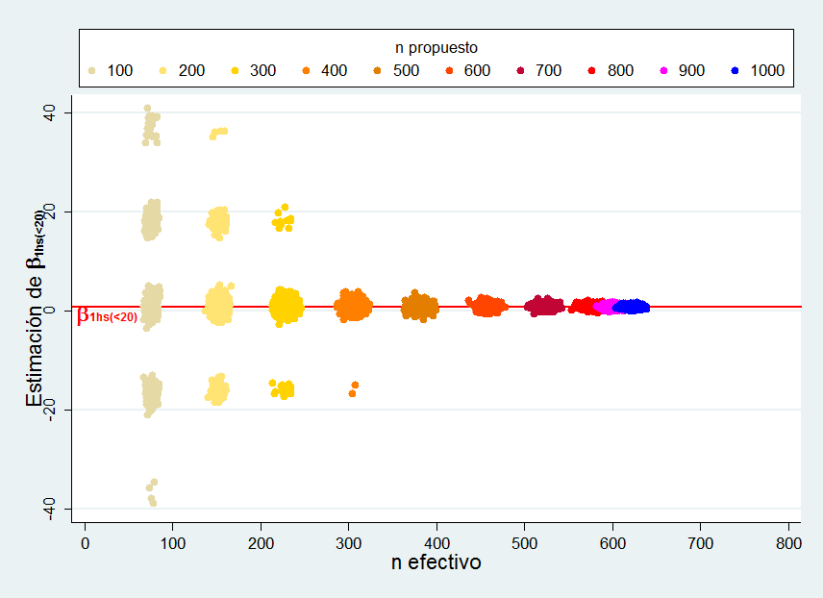
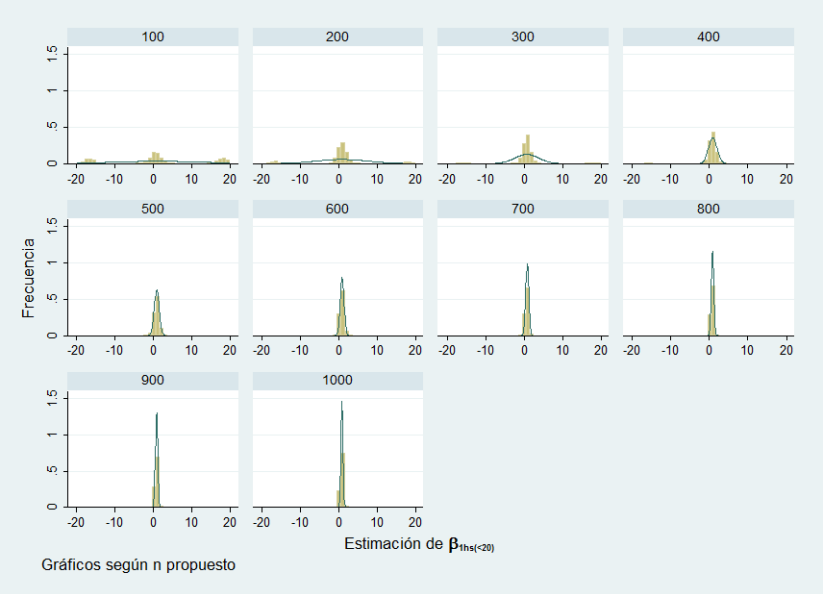


Figura 4.8: Distribución de $\hat{\beta}_{1hs(<20)}$ bajo muestreo por conglomerados para diferentes tamaños de muestra



En lo que refiere a los errores estándares de las estimaciones, se encontró que no se pudieron obtener en algunas de las muestras seleccionadas. Esto se debe a que no se puede estimar la matriz de covarianzas en aquellas muestras donde la proporción de unos es muy pequeña o cero, en alguna de las indicatrices de las covariables.

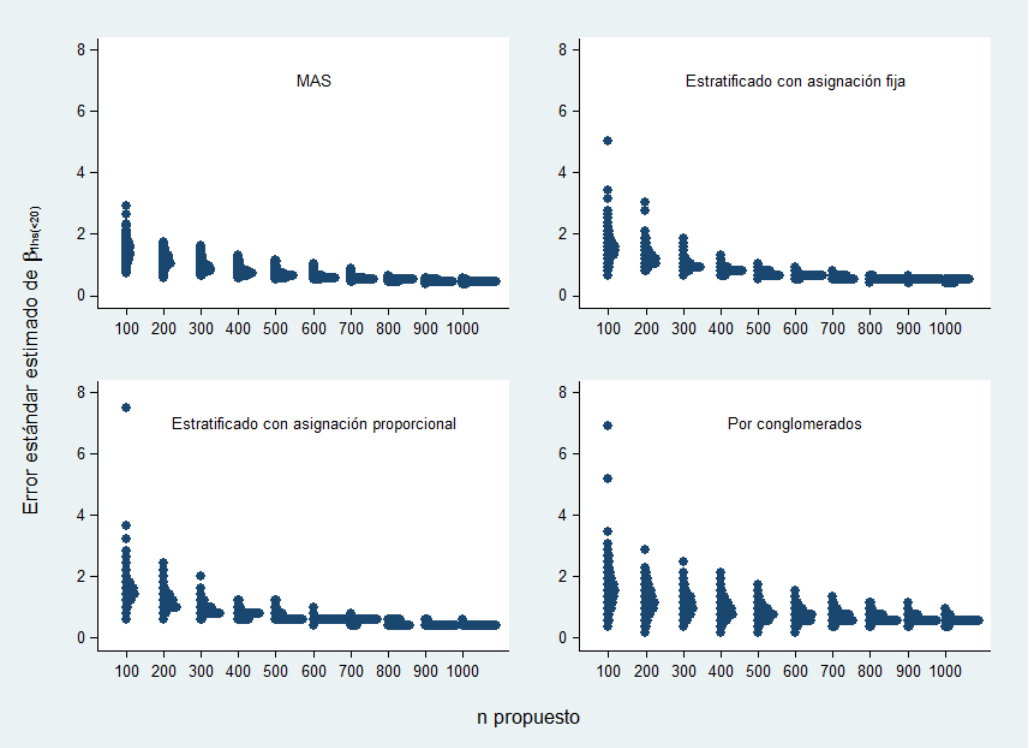
En la Tabla 4.12 se observa que para muestras de tamaño igual o menor a 400 hubo un porcentaje de las 1000 muestras seleccionadas en que no se pudo estimar los errores de las estimaciones. El porcentaje mayor ocurre dentro del muestreo por conglomerados (21 %), seguido por el muestreo estratificado con asignación proporcional (16 %), el muestreo aleatorio simple (14 %) y el muestreo estratificado de asignación fija (11 %), todos para un tamaño propuesto de 100 observaciones. Para n propuesto igual a 200, el porcentaje de muestras donde no se pudo estimar los errores es considerablemente menor, variando entre 1 % (muestreo estratificado con asignación fija) y 3 % (muestreo por conglomerados). Considerando $n = 300$, el porcentaje en cuestión es menor a uno para todos los tipos de muestreo ($0{,}3 - 0{,}5$ %); y para $n = 400$ es cero para conglomerados y estratificado con asignación fija, 0,1 % para MAS y 0,2 % para estratificado con asignación proporcional.

Tabla 4.12: Porcentaje de las 1000 muestras en las que no se pudo estimar el error estándar por tipo de muestreo y tamaño de muestra propuesto

n propuesto	Muestreo			
	MAS	Estratificado según asignación		Conglomerados
		Fija	Proporcional	
100	13,9	11,0	15,6	21,1
200	1,7	0,7	2,8	3,4
300	0,3	0,3	0,5	0,4
400	0,1	0,0	0,2	0,0

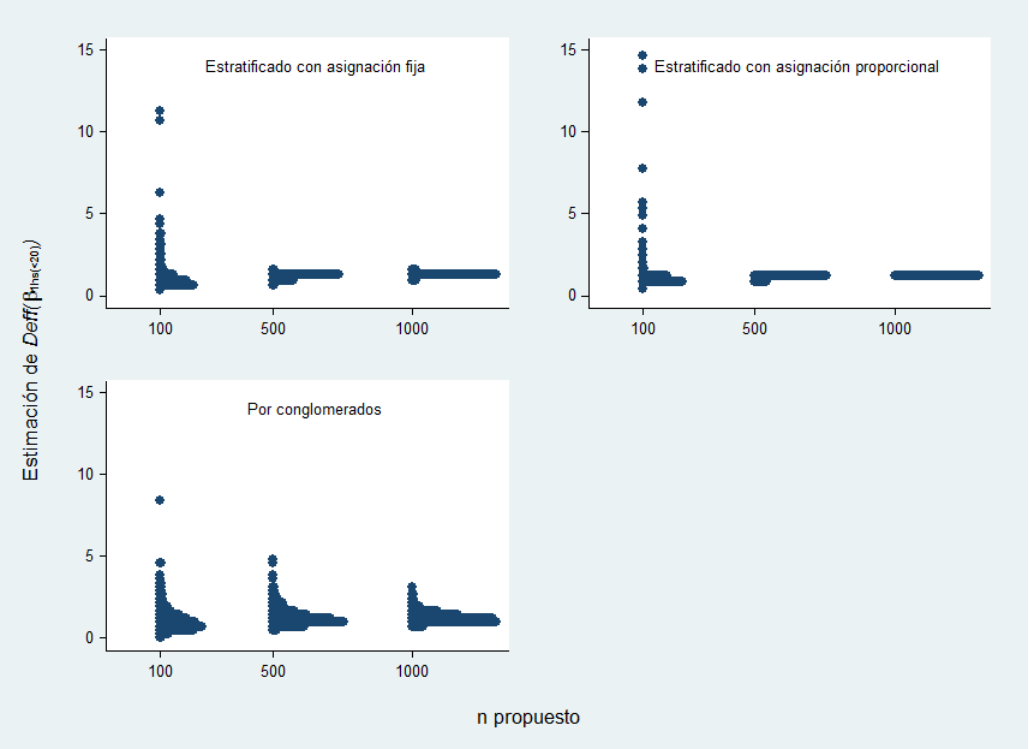
Por lo tanto, el análisis de las estimaciones de los errores estándares de los coeficientes se realizó sin tener en cuenta las muestras donde no se pudieron calcular los errores. En la Figura 4.9 se observa que los errores son más pequeños en el MAS que en los otros tipos de muestreo para las muestras con n propuesto menor o igual a 300. Además, se visualiza que a medida que aumenta el tamaño de muestra el error decrece. De los gráficos se destaca que en los muestreos estratificados y por conglomerados, con $n = 100$, aparecen muestras donde la estimación del error de $\hat{\beta}_{1hs(<20)}$ es "grande". Asimismo, la estabilidad de los errores ocurre a partir de 800 observaciones para MAS y estratificado con asignación fija, y a partir de $n = 900$ para estratificado con asignación proporcional. Sin embargo, en el muestreo por conglomerados el error de $\hat{\beta}_{1hs(<20)}$ decrece a medida que aumenta el tamaño de muestra, pero no alcanza a estabilizarse.

Figura 4.9: $\hat{s}(\hat{\beta}_{1hs(<20)})$ bajo diferentes tamaños de muestra según tipo de muestreo



Como análisis complementario al comportamiento de los errores, se estudió el efecto del diseño para los diferentes tipo de muestreo y para n propuesto igual a 100, 500 y 1000. A partir de los resultados se encontraron dos casos de valores atípicos: $\hat{deff}(\hat{\beta}_{1hs(<20)}) = 26$ en muestreo estratificado con asignación fija y $\hat{deff}(\hat{\beta}_{1hs(<20)}) = 66$ en muestreo estratificado con asignación proporcional. En la Figura 4.10 se presentan los resultados de la estimación del efecto de diseño sin tener en cuenta estos casos atípicos. El *deff* se comporta de manera similar a los errores analizados en el párrafo anterior. Por un lado, en las muestras de tamaño 100, se encontraron casos donde el efecto de diseño es de gran magnitud, por encima de cinco; y por otro, para n=500 y n=1000 el *deff* se estabiliza en los muestreos estratificados, pero la consistencia no sucede en el muestreo por conglomerados.

Figura 4.10: $\hat{deff}(\hat{\beta}_{1hs(<20)})$ bajo diferentes tamaños de muestra según tipo de muestreo



4.2.3. Bondad de ajuste

El ajuste del modelo con toda la población fue evaluado utilizando el estadístico propuesto por Fagerland *et al.* (2008), para $g = 3, 4, 5, 6$, tal como se mostró en la sección 4.1. En este apartado se muestra el desempeño de este estadístico bajo los diferentes planes de muestreo, empleando la misma cantidad de grupos.

La Tabla 4.13 presenta el porcentaje de las 1000 muestras en que no se pudo estimar el estadístico C_g de Fagerland considerando $g = 3, 4, 5, 6$, para cada tipo de muestreo y tamaño de muestra propuesto.

En primer lugar, se destaca que, con $g = 3$ prácticamente no hubo inconvenientes de calcular el estadístico para cualquier tamaño de muestra mediante muestreo estratificado con asignación fija. Sin embargo, C_3 no se pudo calcular en un porcentaje pequeño de las muestras de tamaño 100 y 200 bajo MAS, muestreo estratificado con asignación proporcional y muestreo por conglomerados ($\leq$ 0,06).

Teniendo en cuenta $g = 4$, no se pudo evaluar la bondad de ajuste del modelo en: (a) el 7 % de las muestras MAS, variando entre 6 % ($n = 1000$) y 13 % ($n = 100$); (b) el 1 % de las muestras con estratificado y asignación fija, siendo 0 % para $n = 1000$ y 5 % para $n = 100$; (c) el 7 % con estratificado y asignación proporcional, con valores entre 5 % ($n = 1000$) y 15 % ($n = 100$); y (d) el 2 % de muestreo por conglomerados. Con respecto a esto último, se debe destacar que fue posible calcular el estadístico para todas las muestras de $n \geq 700$ bajo muestreo por conglomerados.

Al evaluar el desempeño del estadístico C_5, este no se pudo obtener en el 25 % y el 44 % de las muestras bajo MAS; entre el 10 % y 31 % de las muestras con muestreo estratificado y asignación fija; entre el 22 % y 41 % con muestreo estratificado con asignación proporcional; y entre el 9 % y 35 % mediante muestreo por conglomerados; para $n = 100$ y $n = 1000$, respectivamente.

Por último, se encontró que el porcentaje de muestras en las que no se puede calcular el estadístico de Fagerland es de gran magnitud si se considera $g = 6$: 65 % de las muestras obtenidas con MAS y estratificado con asignación proporcional, 50 % con muestreo estratificado con asignación fija y 46 % con muestreo por conglomerados.

Tabla 4.13: Porcentaje de las 1000 muestras en que no se pudo calcular el estadístico C_g por tipo de muestreo y tamaño de muestra propuesto

Muestreo	n propuesto	Estadístico de Fagerland *et al.* (2008)			
		C_6	C_5	C_4	C_3
MAS	100	77,8	44,4	12,8	0,6
	200	70,4	34,0	8,3	0,1
	300	66,4	32,2	6,6	0,0
	400	67,1	30,8	8,1	0,0
	500	64,8	32,2	6,3	0,0
	600	63,6	29,3	6,4	0,0
	700	59,8	28,1	6,2	0,0
	800	61,9	28,9	5,8	0,0
	900	60,6	27,2	6,8	0,0
	1000	58,7	25,2	5,5	0,0
	Total	64,9	31,0	7,2	0,1
Estratificado con asignación fija	100	65,1	31,3	5,1	0,0
	200	59,6	26,2	2,6	0,0
	300	58,8	23,3	2,0	0,0
	400	52,0	19,3	1,4	0,0
	500	53,0	15,8	1,1	0,0
	600	46,9	14,3	0,7	0,0
	700	45,1	14,9	0,2	0,0
	800	40,2	12,9	0,3	0,0
	900	38,4	9,8	0,5	0,0
	1000	35,0	11,1	0,0	0,0
	Total	49,4	17,9	1,4	0,0
Estratificado con asignación proporcional	100	74,4	40,7	14,5	0,6
	200	70,5	39,4	9,3	0,2
	300	69,9	38,2	7,3	0,0
	400	66,0	33,4	6,7	0,0
	500	66,1	33,1	4,3	0,0
	600	63,7	28,3	6,3	0,0
	700	63,8	27,2	5,8	0,0
	800	60,5	26,0	5,1	0,0
	900	58,4	24,7	4,8	0,0
	1000	60,7	22,2	5,1	0,0
	Total	65,2	31,2	6,8	0,1
Por conglomerados	100	67,7	34,6	7,2	0,1
	200	60,2	28,9	5,3	0,0
	300	55,3	25,2	2,8	0,0
	400	52,2	25,1	0,7	0,0
	500	48,5	21,3	0,5	0,0
	600	46,4	19,4	0,1	0,0
	700	41,5	20,8	0,0	0,0
	800	37,0	16,6	0,0	0,0
	900	27,9	13,7	0,0	0,0
	1000	29,4	9,2	0,0	0,0
	Total	46,1	21,1	1,5	0,0

Capítulo 5

Conclusiones

En este trabajo se analizó el modelo *logit* multinomial considerando diferentes esquemas de muestreo y se arribó a las siguientes conclusiones.

Primero, si se va a realizar una encuesta con el fin de ajustar un modelo *logit* multinomial, a la hora de calcular el tamaño muestral necesario se debe tener en cuenta la pérdida de observaciones en alguna de las posibles covariables. Es decir, si se incorpora al modelo una pregunta de la encuesta que no corresponde realizar a toda la muestra, el modelo será estimado con la cantidad de casos que corresponde a la pregunta y no con el total de observaciones. Aquí, los resultados mostraron que la pérdida de observaciones es mayor en el muestreo por conglomerados (con igual cantidad de observaciones en cada conglomerado) que en el muestreo aleatorio simple y el muestreo estratificado.

Segundo, las estimaciones de los parámetros poblacionales del modelo resultan extremadamente sesgadas para muestras de 400 o menos observaciones en los cuatro tipos de muestreo considerados. Este sesgo no tiene una dirección definida sino que ocurren tanto subestimaciones como sobreestimaciones. En el análisis se encontró que la dirección del sesgo depende de cuántas observaciones se tenga en cada una de las categorías de las variables regresoras.

Hay que mencionar, además, que la distribución Normal de las estimaciones de los coeficientes se cumple para muestras de al menos 800 unidades, en el caso del muestreo aleatorio simple y el muestreo estratificado, y para muestras de 900 unidades o más en el muestreo por conglomerados.

Tercero, al estudiar la estimación del error estándar de las estimaciones, sucede que en muestras de tamaño pequeño ($n \leq 300$) existe la posibilidad de que no se puedan estimar

los errores. Según el diseño de muestreo propuesto, esto ocurre con mayor frecuencia en el muestreo por conglomerados, seguido por el muestreo estratificado con asignación proporcional, el muestreo aleatorio simple y por último, el muestreo estratificado con asignación fija, y sucede debido a la no existencia de observaciones en alguna de las categorías de las covariables. A su vez, los errores convergen para muestras con 900 observaciones o más y extraídas mediante muestreo aleatorio simple o estratificado, pero no convergen en el muestreo por conglomerados aquí utilizado.

Dicho lo anterior, y teniendo en cuenta que en la práctica las encuestas requieren de muestreos más complejos que los aquí empleados, el tamaño de muestra mínimo necesario para especificar un modelo *logit* multinomial es de al menos 900 observaciones. Una posible continuación de este trabajo es determinar cuál es el tamaño de muestra adecuado si se trabaja con muestreos más complejos, que impliquen varios tipos de muestreo y varias etapas.

Finalmente, el desempeño del test de bondad de ajuste para regresión *logit* multinomial propuesto por Fagerland *et al.* (2008) es inestable según la cantidad de grupos que se utilice. Esto puede estar asociado a lo que los autores del estadístico plantean como limitaciones del trabajo: por un lado, las simulaciones del test C_g fueron realizadas utilizando una variable regresora continua, tres categorías para la variable de respuesta y dos tamaños de muestra ($n = 100$ y $n = 400$); y por otro, plantean que no han estudiando la elección ideal del número g.

Bibliografía

[1] Boado, M. (2005) *Una aproximación a la deserción estudiantil universitaria en Uruguay*. Caracas: IESALC/UNESCO.

[2] Fagerland, M. W., Hosmer, D. W., y Bofin, A. M. (2008). *Multinomial goodness-of-fit tests for logistic regression models*. Statistics in Medicine, 27(21), 4238-4253.

[3] Fagerland, M. W., and Hosmer, D. W. (2012). *A generalized HosmerLemeshow goodness-of-fit test for multinomial logistic regression models*. Stata Journal 12, 447–453.

[4] Heeringa, S.G., West, B.T. y Berglund, P.A. (2010). *Applied Survey Data Analysis*. Boca Raton, London and New York: CRC Press.

[5] Hosmer, D.W. Jr., Lemeshow, S. y Sturdivant, R.X. (2013). *Applied Logistic Regression, Third Edition*. Wiley Series in Probability and Statistics. New Jersey: John Wiley and Sons Inc.

[6] Kalton, G. (1983). *Introduction to Survey Sampling*. Sage University Paper series on Quantitative Application in the Social Sciences, 07-035.

[7] Lehtonen, R. y Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys, 2nd Ed*. Chichester: John Wiley & Sons.

[8] Levy, P. y Lemeshow, S. (2008). *Sampling of Populations: Methods and Applications, 4th Ed*. New York: John Wiley & Sons.

[9] Lohr, S. (2009). *Sampling: Design and Analysis, 2nd Ed*. Boston: Cengage Learning.

[10] Ministerio Secretaría General de la Presidencia de Chile (2008). Sobre acceso a la información pública. En *Biblioteca del Congreso Nacional de Chile* [En línea]. Disponible en: `http://bcn.cl/1uuq2`. Visitado el 30 de mayo de 2016.

[11] Poder Legislativo de Uruguay (2008). Derecho de acceso a la información pública. En *Parlamento del Uruguay* [En línea]. Disponible en: `http://www.parlamento.gub.uy/leyes/AccesoTextoLey.asp?Ley=18381&Anchor=`. Visitado el 30 de mayo de 2016

[12] Särndal, C.E., Swensson, B. y Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer series in Statistics. New York: Springer-Verlag.

[13] StataCorp(2013). *Stata: Release 13.* Statistical Software. College Station, TX: StataCorp LP.

[14] Universidad de la República (1999a). *Plan de Estudios 2000 de la Facultad de Química.* Aprobado por el Consejo Directivo Central de la Universidad de la República el 23 de noviembre de 1999.

[15] Universidad de la República (1999b). *Plan de Estudios 2000 para la carrera de Ingeniería Química.* Aprobado por el Consejo Directivo Central de la Universidad de la República el 23 de noviembre de 1999.

[16] Universidad de la República (2003a). *Exigencias académicas para acceder al título de Licenciado en Química.* Aprobado por el Consejo Directivo Central de la Universidad de la República el 9 de noviembre de 2003.

[17] Universidad de la República (2003b). *Plan de Estudios de la carrera de Ingeniería de Alimentos.* Aprobado por el Consejo Directivo Central de la Universidad de la República el 12 de agosto de 2003.

[18] Wolter, K. (2007). *Introduction to variance estimation, 2nd Ed.* New York: Springer.

[19] Zeng, Q. (2011). *Developing Sampling Weights for Complex Surveys. An Approach to the School Physical Activity and Nutrition (SPAN) Project.* Saarbrücken: Lap Lambert Academic Publishing GmbH & Co. KG