MATEMÁTICA
FACES, año 9, N° 17, mayo/agosto de 2003.

Análisis de encuestas basado en diseño y modelos muestrales[1]

Analyzing surveys by using design-based And model-based inferential methods

Natacha Liseras[2]

RESUMEN / SUMMARY

En el presente trabajo se comparan dos métodos de análisis de encuestas: inferencia basada en diseño muestral (inferencia clásica) e inferencia basada en modelos, para el caso específico de datos binarios correlacionados. Se propone la formulación de modelos marginales y de modelos mixtos basados en la función de verosimilitud completa. Se efectúa una aplicación concreta a un estudio de corte transversal en el cual la dependencia entre las observaciones se debe al submuestreo de unidades primarias alumnos encuestados dentro de facultades. La variable respuesta es la presencia de vocación emprendedora en alumnos universitarios de economía, administración e ingeniería, y se estima la proporción de alumnos con vocación emprendedora en 0.4 bajo ambos métodos. El estudio realizado permite concluir que la inferencia basada en modelos otorga mayor flexibilidad de análisis que la inferencia clásica basada en diseño muestral.

[1]Este trabajo constituye una síntesis de la tesis de maestría en Estadística Aplicada en la Universidad Nacional de Córdoba.

La versión completa disponible en: http://eco.mdp.edu.ar/cendocu/opac/tesis.htm.

Director de tesis: Dr. Raúl Macchiavelli (Universidad de Puerto Rico).

[2]nliseras@mdp.edu.ar

In this paper, two methods for survey analysis are compared: the design based and the model based inferential methods, for survey analysis of correlated binary data. The formulation of marginal generalized linear models and random-effect models are proposed based on a logic link. A concrete application on a cross-sectional study is carried out, in which dependence among data arises because of sub sampling of primary units students surveyed in their schools. The variable response shows an entrepreneurial vocation among undergraduates of Economics, Business Administration and Engineering. The estimated rate of students with entrepreneurial vocation is 0.4, to conclude with the fact that the model based estimation renders more analysis flexibility than the classical design based one.

PALABRAS CLAVE / *KEYWORDS*

Observaciones binarias correlacionadas, inferencia basada en modelos, ecuaciones de estimación generalizadas, modelos marginales, modelos de efectos aleatorios, vocación emprendedora

Correlated binary data, model-based inference, generalized estimating equations, marginal models, random-effect models, entrepreneurial vocation

INTRODUCCIÓN

La inferencia estadística permite estimar la proporción de individuos que presentan una característica determinada. La inferencia clásica, basada en el diseño de muestreo, requiere una estricta aleatoriedad en la recolección de los datos y la existencia de buenos marcos muestrales. En la práctica, debido a las falencias de las estadísticas disponibles, es usual que los marcos de información se encuentren incompletos o desactualizados, lo cual plantea un conflicto entre los objetivos propuestos por la investigación y los resultados obtenidos. Lamentablemente, la construcción de marcos apropiados por parte del investigador representa un alto costo en tiempo y dinero que no siempre es posible afrontar.

La inferencia basada en modelos puede realizarse aun si la selección de las unidades de muestreo no se efectúa al azar, cuando puede conceptualizarse el comportamiento de la respuesta como aleatorio. En otras palabras, si la muestra seleccionada puede pensarse como una realización del mismo

modelo de probabilidad que se hubiera generado por un proceso de selección aleatorio. De este modo, la información recolectada se utiliza para construir modelos a partir de los cuales es posible inferir sobre características de la población objetivo.

En general, bajo esta estrategia se usan modelos lineales o modelos lineales generalizados (MLGs). Los primeros asumen que: *(a)* la varianza de las observaciones es constante; *(b)* la media y la varianza son funcionalmente independientes; *(c)* las perturbaciones aleatorias siguen una distribución normal. Cuando estos supuestos propios del modelo lineal clásico no se cumplen, es posible emplear los MLGs.

Dichos modelos son especificados mediante la distribución de probabilidad de las observaciones y de una función de enlace que relaciona los parámetros del modelo con la media de la distribución (McCullagh & Nelder, 1989). Los MLGs flexibilizan dos supuestos del modelo lineal clásico, ya que: *(a)* la distribución de las perturbaciones aleatorias puede provenir de una familia exponencial uniparamétrica distinta de la normal; *(b)* el enlace puede ser cualquier función conocida, monótona y diferenciable, sin ser necesariamente la función identidad.

Respecto de las covariables seleccionadas, éstas pueden representar tanto efectos fijos como aleatorios. Si los efectos son fijos, el espacio de inferencia queda limitado a los niveles de los factores que se manifiestan en los datos y el interés reside en estimar los parámetros asociados. Si son aleatorios, el espacio de inferencia consiste en la población de niveles, no todos los cuales se observan. Los modelos lineales generalizados mixtos resultan adecuados si se contempla la inclusión de términos fijos y aleatorios en el predictor lineal.

En principio, tanto los modelos lineales como los MLGs suponen que las observaciones son independientes. Cuando se desea modelar respuestas dependientes, por ejemplo debido al submuestreo al interior de *clusters*, es necesario utilizar distintas extensiones de los modelos lineales generalizados, pudiendo optarse por los modelos marginales o los modelos mixtos. Su uso representa un importante aporte metodológico para las Ciencias Sociales, dado que su campo usual de aplicación es el de las Ciencias Biológicas.

Ignorar la correlación, generalmente, conlleva la subestimación de los errores estándares de los parámetros y la falta de precisión en la inferencia; además, la varianza de los parámetros se estima en forma inconsistente.

Contemplar en qué medida las observaciones pertenecientes a un mismo grupo son dependientes entre sí: *(a)* hace posible una mejor estimación de los efectos fijos del modelo y sus errores estándares; *(b)* permite conocer la influencia que ejercen los efectos a nivel *cluster* sobre el comportamiento individual y la manera en que dichos efectos operan (Rodríguez & Goldman, 1995).

Con un modelo marginal se describe la esperanza de la variable respuesta como función de covariables y se especifica una estructura de dependencia entre las observaciones, dirigiendo la inferencia hacia el promedio de la población *(population average inference)*. Según Zeger & Liang (1986), el enfoque marginal resulta adecuado si el interés recae en estimar los parámetros asociados a las covariables para la esperanza marginal y la correlación entre las observaciones es considerada como ruido. Si estimar los coeficientes para cada *cluster* fuese relevante, debería optarse por los modelos lineales generalizados mixtos.

Con un modelo mixto, la inferencia es específica para cada *cluster (cluster specific inference)*. Si se formula un modelo mixto con verosimilitud completa, se describe la esperanza de la variable respuesta condicional a parámetros aleatorios específicos para cada *cluster*, lo que requiere establecer un supuesto acerca de la distribución de probabilidad de dichos efectos aleatorios, la que típicamente se supone normal (Diggle *et al.*, 2002).

Los parámetros estimados representan los efectos de las covariables sobre las chances de un *cluster* particular. Por ejemplo, los coeficientes de regresión describen la respuesta de cada *cluster* ante cambios en el nivel de las covariables, estimándose el cambio esperado en las probabilidades individuales (Zeger *et al.*, 1988).

El presente trabajo se propone efectuar una comparación entre dos métodos de inferencia: basada en diseños de muestreo y basada en modelos. Ambas estrategias de análisis se aplican a un caso concreto, que consiste en estimar la proporción de alumnos universitarios con vocación emprendedora. La dependencia entre observaciones binarias se debe a que las respuestas provienen de un muestreo por conglomerados en dos etapas.

El interés por estudiar la vocación emprendedora surge a partir de la reconocida trascendencia que el proceso de creación de empresas tiene sobre el desarrollo económico de una región. El caso de los alumnos universitarios

es de particular importancia, ya que se trata de individuos capaces de emprender proyectos innovadores y de gestión profesionalizada.

Objetivos

Los principales objetivos de este trabajo son:

- Comparar la aplicación de métodos de inferencia basados en diseño muestral y en modelos, destacando ventajas y desventajas de cada uno de ellos.
- Estimar la proporción de alumnos universitarios con vocación emprendedora en carreras de economía, administración e ingeniería de facultades públicas y privadas de la Ciudad Autónoma de Buenos Aires y de la Provincia de Buenos Aires.
- Aplicar metodologías modernas de modelación en el área de las Ciencias Sociales.

Aplicación: Población objetivo y diseño muestral

La población objetivo del estudio está formada por los alumnos que cursan el último año de carreras de economía, administración e ingeniería en facultades públicas y privadas de la Ciudad Autónoma de Buenos Aires y de la Provincia de Buenos Aires (República Argentina). El marco de información utilizado consiste en una lista de las titulaciones dictadas en cada una de las facultades del área de cobertura de la investigación, junto con el número de alumnos inscriptos en el último año de las titulaciones seleccionadas. Las encuestas fueron realizadas en noviembre de 2002.

Al tener en cuenta la carrera, el carácter público o privado y la localización geográfica de las instituciones, se conforman ocho estratos, en cada uno de los cuales se han muestreado dos facultades al azar. La muestra de alumnos (elemento de muestreo) se ha obtenido seleccionando al azar los cursos (unidad de muestreo) en los cuales efectuar las encuestas. Esta es una técnica de muestreo utilizada por cuestiones operativas, ya que en la práctica no es factible disponer del listado de alumnos que cursan el último año para seleccionar un subconjunto de ellos, sin que la selección de cursos sea considerada como una etapa adicional. Por consiguiente, el diseño consta de dos etapas: la selección al azar de facultades y el relevamiento de todos los alumnos dentro de los cursos seleccionados. Este puede considerarse equiva-

lente a la selección al azar de alumnos dentro de cada facultad.

En la Tabla 1 pueden observarse las características de las facultades muestreadas en cuanto a carrera dictada, tipo de gestión y localización geográfica de la misma. Esta clasificación reviste interés para la interpretación de los resultados.

DEFINICIÓN DE LA VARIABLE RESPUESTA

La variable respuesta es la presencia de vocación emprendedora en el alumno, al ser esta de naturaleza binaria. Surge en forma objetiva de las encuestas realizadas, en las cuales se plantean tres alternativas: *(a)* que el alumno haya creado alguna vez una empresa; *(b)* que posea una idea concreta de negocios sin haber creado su propia empresa; *(c)* que opine que al graduarse le gustaría crear una empresa, pero al momento no posea ninguna idea de negocios ni haya creado una empresa.

Se define que un alumno posee vocación emprendedora (VE=1) si responde afirmativamente a las alternativas *(a)* y *(b)* y que no posee vocación emprendedora (VE=0) si responde negativamente a los tres ítems. A fin de poder captar más claramente la presencia de vocación emprendedora, se excluyen del análisis 149 alumnos que responden de acuerdo a la alternativa *(c)*, dado que ellos no pertenecen claramente a ninguno de ambos grupos. De este modo, la muestra queda compuesta por 799 alumnos, ajustándose el modelo con las 723 encuestas disponibles sin datos faltantes.

INFERENCIA BASADA EN DISEÑO MUESTRAL (INFERENCIA CLÁSICA)

En esta sección se presentan los resultados obtenidos al aplicar la inferencia basada en diseño muestral para hallar estimadores puntuales y por intervalo de la proporción de alumnos universitarios con vocación emprendedora (VE). Se estima esta proporción para la totalidad de la población objetivo y, luego, se obtienen estimaciones para distintos dominios o subpoblaciones, analizando si existen diferencias estadísticamente significativas entre ellas.

Las fórmulas aplicadas corresponden al muestreo por conglomerados en dos etapas. Si los individuos pertenecientes a los conglomerados *(clusters)* se muestrean en lugar de enumerarse completamente, el diseño es multietápico. Un muestreo en dos etapas consiste en seleccionar primero una muestra

aleatoria de conglomerados y, posteriormente, una muestra aleatoria de los elementos que ellos contienen.

Para estimar la media global, o proporción en la población, no existe una única fórmula a utilizar bajo este tipo de diseño. Es posible optar por los estimadores de proporciones o los estimadores de razón si se considera que el tamaño del *cluster* es una variable aleatoria. Esto conlleva la dificultad de la estimación de la varianza y, como consecuencia, de la amplitud de los intervalos de confianza, que también dependen de la elección efectuada.

Mediante el uso de los estimadores para proporciones bajo un muestreo por conglomerados en dos etapas, la media global se estima como (Scheaffer *et al.*, 1987):

$$\hat{\mu}_{..} = \frac{\sum_{i=1}^{k} M_i \hat{\mu}_{i.}}{\sum_{i=1}^{k} M_i}$$

Como M_i representa el número de elementos en el i-ésimo *cluster*, el estimador le da un peso mayor a los conglomerados más grandes. Esta expresión supone que se desconoce la cantidad de individuos en la población, por lo cual $\overline{M}$ se calcula con información proveniente de la muestra. La varianza estimada de la media global es (Scheaffer *et al.*, 1987):

$$Var\left(\hat{\mu}_{..}\right) = \left(\frac{K-k}{K}\right)\left(\frac{1}{k\overline{M}^2}\right)S_1^2 + \frac{1}{kK\overline{M}^2}\sum_{i=1}^{k} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\left(\frac{\hat{\mu}_{i.}\left(1-\hat{\mu}_{i.}\right)}{m_i - 1}\right) =$$

$$= \left(1-f_1\right)\left(\frac{1}{k\overline{M}^2}\right)S_1^2 + \frac{1}{kK\overline{M}^2}\sum_{i=1}^{k} M_i^2\left(1-f_{2i}\right)\left(\frac{\hat{\mu}_{i.}\left(1-\hat{\mu}_{i.}\right)}{m_i - 1}\right)$$

$$S_1^2 = \frac{\sum_{i=1}^{k} M_i^2\left(\hat{\mu}_{i.} - \hat{\mu}_{..}\right)^2}{k-1}$$

Siendo:

1 $\overline{M}$ el tamaño promedio del *cluster*.
2 k el número de *clusters* en la muestra.

3 K el número de *clusters* en la población.

4 f_1 la fracción de muestreo de primera etapa.

5 f_{2i} la fracción de muestreo de segunda etapa, distinta para cada *cluster*.

6 S_1^2 la varianza entre las medias de *cluster*.

Para comparar si las proporciones en dos dominios $\mu_{..}^{(1)}$ y $\mu_{..}^{(2)}$ difieren entre sí, bajo la hipótesis nula que establece que ambas proporciones son iguales, el estadístico de prueba posee distribución normal estándar para grandes muestras. Éste se define como:

$$z = \frac{\hat{\mu}_{..}^{(1)} - \hat{\mu}_{..}^{(2)}}{\sqrt{Var\left(\hat{\mu}_{..}^{(1)}\right) + Var\left(\hat{\mu}_{..}^{(2)}\right)}} \sim N(0,1)$$

El estimador puntual de la media global $(\hat{\mu}_{..})$ es 0.40, valor que representa la proporción de alumnos universitarios con VE en la población objetivo Los valores estimados se resumen en la Tabla 2.

Al realizar un análisis por tipo de gestión de la facultad, en la Tabla 3 se incluyen las medias y varianzas estimadas para cada uno de los dominios, observándose una mayor vocación emprendedora entre los alumnos de instituciones privadas (valor-p < 0.001). En la Tabla 4 se reportan las proporciones estimadas por carrera, sin que existan diferencias estadísticamente significativas entre ellas (valor-p = 0.56).

Asimismo, es de interés calcular las proporciones de alumnos con vocación emprendedora por género. En este caso se presenta un problema debido a que no existe información disponible acerca de la cantidad de hombres y mujeres que cursan las carreras bajo análisis. Por tal motivo, se establece como supuesto que los porcentajes por género para cada carrera son los que surgen de la muestra (Economía y administración: 55% mujeres y 45% hombres; Ingeniería: 10% mujeres y 90% hombres). En la Tabla 5 se indican la media global y la varianza para cada género calculadas con este supuesto, siendo la proporción de hombres con VE mayor que la proporción de mujeres con VE (valor-p < 0.001).

Para concluir, vale destacar algunas limitaciones que surgen de la aplicación de la inferencia clásica:

1 Estimar correctamente la media y la varianza requiere conocer el

Tamaño total del *cluster* del que se extrae la muestra. Luego, las variables de clasificación a utilizar para definir las subpoblaciones deben ser aquéllas para las cuales existe información suficiente.

2 Cuando se desean comparar las proporciones estimadas para dos subpoblaciones, es necesario particionar la muestra, calculándose la media global y su varianza con un distinto número de observaciones en cada subpoblación. Tal como puede apreciarse en las Tablas 2 a 5, a menor tamaño de muestra, más imprecisa es la estimación.

OBSERVACIONES CORRELACIONADAS

Al menos tres razones permiten pensar que los alumnos que concurren a una misma facultad son similares entre sí en numerosos aspectos difíciles o imposibles de medir. En primer lugar, ellos deciden a qué institución asistir, y es sensato pensar que individuos que eligen la misma facultad se asemejan e.g., grupo social de pertenencia. En segundo lugar, existen variables a nivel facultad que no pueden aislarse y que afectan a todos los alumnos en forma simultánea e.g., modalidad de dictado de las materias. Por último, los individuos dentro de una misma facultad, particularmente tratándose de estudiantes a punto de graduarse, interactúan y ejercen influencia unos sobre otros, creando redes personales en las cuales circula información de distinto tipo.

Las características mencionadas explican por qué es válido considerar que las observaciones no son independientes entre sí. Ello habilita, por consiguiente, a emplear métodos de análisis que contemplan la dependencia existente entre las mediciones en una misma facultad. En los modelos formulados, la presencia de vocación emprendedora es descripta como función de efectos tanto fijos como aleatorios.

En la Tabla 6 se incluyen las covariables seleccionadas junto con su codificación y la respectiva hipótesis de trabajo, que son: GÉNERO, OCUPADO, ACTITUD, VISIÓN, RIESGO y CREATIVIDAD, todas ellas binarias. Dada la parametrización adoptada, los estimadores se interpretarán como el cociente entre las chances de que un alumno posea vocación emprendedora cuando cada covariable, controlada por las restantes covariables, valga 1 respecto de cuando valga 0.

INFERENCIA BASADA EN MODELOS

Modelos formulados

Bajo el modelo marginal, utilizando el enlace *logit*, se describe la media marginal de vocación emprendedora (VE) como función de las covariables y, separadamente, se especifica la estructura de dependencia intra-*cluster*. En esta versión reducida del trabajo sólo se presentan los resultados correspondientes al modelo marginal con estructura de correlación intercambiable o de simetría compuesta:

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = \eta_{ij}$$

$$i = 1,\ldots,k$$

$$j = 1,\ldots,m_i$$

$$\eta_{ij} = \beta_0 + \beta_1\, GENERO_{ij} + \beta_2\, OCUPADO_{ij} + \beta_3\, ACTITUD_{ij} + \beta_4\, VISION_{ij} + \\ + \beta_5\, RIESGO_{ij} + \beta_6\, CREATIV_{ij}$$

$$\text{corr}(y_{ir}, y_{is}) = \alpha \qquad \forall r \neq s$$

El método de ecuaciones de estimación generalizadas *(Generalized Estimating Equations - GEE)* de Zeger & Liang (1986) y Liang & Zeger (1986), extiende el uso de la cuasi-verosimilitud al análisis de observaciones dependientes. Dicho método ofrece la ventaja de ser sencillo computacionalmente y aplicable a una clase general de funciones de distribución y de enlace, al permitir que los *clusters* difieran de tamaño (Longford, 1994; Agresti, 2002).

Tal método incorpora una matriz de correlación propuesta o de trabajo *(working correlation matrix)* dentro de las ecuaciones de estimación, las cuales se resuelven mediante algún algoritmo iterativo. Los parámetros de regresión se obtienen resolviendo un sistema de ecuaciones tipo *score* (SAS Institute Inc., 1999).

Los estimadores *GEE* de los parámetros del modelo son consistentes aun si se especifica incorrectamente la estructura de covarianza. (Spiess & Hamerle, 2000; McCulloch & Searle, 2001; Agresti, 2002). La varianza de los estimadores se estima en forma robusta mediante la denominada matriz "*sandwich*", la cual protege contra la elección de una estructura de correlación incorrecta para los parámetros estimados, al utilizar la evidencia

empírica de correlación para ajustar los errores estándares (Lipsitz *et al.*, 1994; Fahrmeir & Tutz, 2001; Agresti, 2002).

Bajo el modelo mixto con verosimilitud completa, se incluye un término aleatorio en el predictor lineal para representar el efecto de *cluster*, cuya distribución de probabilidad se propone que es normal. El enlace utilizado también es el *logit*:

$$g(\mu_{ij}/U_i) = \text{logit}(\mu_{ij}/U_i) = \eta_{ij}$$

$$i = 1,\ldots k$$

$$j = 1,\ldots m_i$$

$$\eta_{ij} = \beta_0 + \beta_1 GENERO_{ij} + \beta_2 OCUPADO_{ij} + \beta_3 ACTITUD_{ij} + \beta_4 VISION_{ij} + \\ + \beta_5 RIESGO_{ij} + \beta_6 CREATIV_{ij} + U_i$$

$$U_i \sim N(0, \sigma_{ii}^2)$$

Este modelo se estima mediante el método de máxima verosimilitud completa.

Estimaciones

Para la estimación de los modelos se utilizó SAS, versión 8.2 (SAS Institute Inc., Cary, NC, USA). Los procedimientos de estimación fueron, respectivamente, GENMOD y NLMIXED. Se descarta la existencia de multicolinealidad luego de analizar la matriz de correlación, los índices de condición y los cocientes de chances marginales entre las covariables.

Las covariables resultan estadísticamente significativas pruebas de Wald, (Tablas 7 y 8). En los modelos marginales, el cálculo de los estadísticos de cuasi-*score* para pruebas de tipo III, indican que la remoción de alguna covariable del predictor lineal tiene un efecto estadísticamente significativo en el ajuste. En el modelo mixto, se observa que los estadísticos AIC y BIC son menores para el modelo completo.

Aun cuando dos interacciones dobles son estadísticamente significativas, el hecho de que las mismas sean ordenadas por ejemplo, los efectos no se cruzan, hace que cambie la pendiente pero no el sentido de asociación entre las variables. Por tal motivo, como no se modifica la interpretación de los efectos principales, se opta por un modelo más parsimonioso que no incluya

los términos de interacción.

La interpretación de los coeficientes difiere en ambas estrategias de análisis, aun cuando todos los coeficientes se interpretan en términos de cocientes de chances condicionales. Ello implica que las preguntas y objetivos que guían la investigación actúen como determinantes del enfoque estadístico a elegir, independientemente de otras consideraciones.

Los coeficientes estimados bajo el enfoque marginal representan un cociente de las chances promedio en la población e ignoran el efecto que ejerce la facultad particular a la que concurre el alumno, aunque condicionan sobre las otras covariables del modelo, (Tabla 9). Si los parámetros son estimados mediante un modelo mixto, representan cocientes de chances condicionales controlados por el efecto de *cluster*, por ejemplo, correspondientes a una facultad determinada, (Tabla 10).

Probabilidades estimadas

Una vez finalizado el proceso de estimación, es interesante expresar los parámetros estimados en términos de probabilidades. Para ello, a continuación se estima la probabilidad de que un alumno universitario perteneciente a la categoría denominada modal posea vocación emprendedora (VE), para cada uno de los modelos ajustados. Las modalidades de las covariables que componen dicha categoría surgen de la tabla de contingencia parcial a cinco vías de clasificación que es observada con mayor frecuencia. Dicha categoría modal corresponde a un hombre ocupado, que carece de actitud empresarial frente al desempleo aunque valora dicha actividad favorablemente, adverso al riesgo y con un nivel medio-bajo de creatividad.

La probabilidad de que un alumno perteneciente a la categoría modal posea vocación emprendedora se estima en 0.41 con el modelo marginal. Dicha probabilidad se ve incrementada si el alumno tiene una actitud empresarial frente al desempleo, es propenso al riesgo o si su nivel de creatividad es alto. En cambio, se ve disminuida si se trata de una mujer, si el individuo es desocupado o inactivo o si tiene una visión desfavorable de la actividad empresarial, (Tabla 11). Los coeficientes de riesgo relativo se calculan como el cociente entre la probabilidad de VE=1 para la modalidad de cada covariable indicada en la tabla y la correspondiente a la categoría

modal, manteniendo sin cambios las otras covariables.

Para estimar las probabilidades y analizar el riesgo relativo bajo el modelo mixto con verosimilitud completa, es necesario tener presente las particularidades de las distintas facultades muestreadas, las cuales fueron explicitadas en la Tabla 1. En este caso, se adiciona al predictor lineal el efecto aleatorio predicho para cada facultad.

En Tabla 12 se indican los cocientes entre las probabilidades estimadas para un alumno perteneciente a la categoría modal de cada una de las facultades y la probabilidad asociada a la categoría modal considerando que el efecto aleatorio es nulo facultad típica. La probabilidad de que un alumno perteneciente a la categoría modal posea vocación emprendedora es de 0.42 en una facultad típica.

Los riesgos relativos mayores a 1 indican que la probabilidad en esa facultad es mayor que en una facultad típica. En la Tabla 12 se observa claramente el impacto positivo que ejerce sobre la probabilidad de poseer vocación emprendedora la gestión privada de la facultad, tal como se demostrara aplicando inferencia clásica.

Validación del modelo

Para evaluar la calidad del ajuste de los modelos formulados, existen pocas herramientas disponibles debido a que: *(a)* la naturaleza binaria de la variable respuesta y de las covariables restringe las técnicas formales y gráficas a aplicar, y *(b)* aún no se han desarrollado técnicas formales adecuadas a los modelos marginales basados en la función de cuasi-verosimilitud. Sin embargo, existen distintas formas de evaluar el poder predictivo de un modelo. A continuación se presentan las curvas ROC, mientras que en la versión completa de este trabajo se calculan también la correlación entre los valores ajustados y observados, la tasa de error aparente y tasa de error por validación cruzada *leave-one-out*.

Las curvas ROC *(Receiver Operating Characteristics)* son gráficos que sintetizan la relación entre la sensibilidad por ejemplo, la probabilidad de que el modelo clasifique al alumno con VE=1, dado que el alumno posee VE y uno menos la especificidad la especificidad es la probabilidad de que el modelo clasifique al alumno con VE=0, dado que el alumno no posee VE para todos los posibles puntos de corte, motivo por el cual resultan muy informativas. Estas

curvas, que conectan los puntos (0,0) y (1,1), son usualmente cóncavas y cuanto mayor es el área debajo de las mismas, mejor es la capacidad de predicción del modelo (Agresti, 2002).

La Figura 1 muestra las curvas correspondientes a los modelos marginal y mixto con verosimilitud completa. Las áreas debajo de las curvas se estiman en 0.7966 y 0.8158 respectivamente, lo cual confiere a los modelos un alto poder de predicción.

COMPARACIÓN ENTRE MÉTODOS DE INFERENCIA

Luego de realizar inferencia clásica e inferencia basada en modelos, resta comparar los resultados hallados por ambas vías. Para que sea factible contrastar los resultados alcanzados por ambos métodos, es necesario estimar la proporción mediante inferencia clásica condicionando respecto de las covariables incluidas en el predictor lineal de los modelos formulados, por ejemplo, estratificando a posteriori. Esta alternativa, en general, tiene como desventajas que el tamaño muestral disminuye drásticamente y que puede aplicarse en tanto las tablas de contingencia parciales contengan información suficiente para estimar la varianza entre los *clusters* y dentro de los mismos, lo cual restringe las posibilidades de análisis. Pero además, en este caso en particular, se suma un problema adicional asociado a los requerimientos de las fórmulas con las que se estiman la media y la varianza: es imposible conocer la cantidad de alumnos en cada facultad que poseen características que surgen a posteriori de las encuestas realizadas.

En la Tabla 13, se presentan las proporciones estimadas de alumnos universitarios de género masculino y femenino con vocación emprendedora bajo ambos métodos, junto con los intervalos de confianza asociados. Para poder obtener resultados comparables mediante el uso de modelos, a GÉNERO se le asigna el valor 1 para estimar la proporción de hombres con VE y el valor 0 para estimar dicha proporción entre las mujeres. Para las demás covariables, la forma de ignorar su influencia es reemplazarlas por las respectivas proporciones muestrales en la subpoblación de hombres y de mujeres.

En la subpoblación de hombres es prácticamente igual de eficiente aplicar inferencia clásica o el enfoque marginal. Para la subpoblación de mujeres, el modelo marginal brinda mayor precisión que la inferencia basada en el diseño muestral. Una cuestión que explica que la estimación con este último método

sea menos precisa para las mujeres que para los hombres es el menor tamaño muestral en el primer grupo.

Las estimaciones obtenidas con el modelo mixto son menos precisas, dado su mayor error estándar. No obstante, los intervalos de confianza son ligeramente más estrechos si, en lugar de referir la inferencia a un *cluster* típico por caso, efecto aleatorio nulo, se adiciona al predictor lineal el promedio ponderado de los efectos aleatorios. La mayor variabilidad que exhiben estos modelos refleja que el espacio de inferencia es más amplio, lo cual está inducido por los niveles aleatorios de la variable FACULTAD. Si el interés reside en conocer el efecto que ejercen las covariables sobre la vocación emprendedora en un *cluster* específico por ejemplo, "controlando" por la facultad a la que asiste el alumno, éste es el único método adecuado a pesar de la menor precisión.

Una segunda apreciación que surge de la Tabla 13, se refiere a las estimaciones puntuales. Las medias estimadas con el uso de modelos no difieren entre sí, pero se ubican por debajo de la proporción estimada mediante la inferencia clásica. La diferencia se explica porque la fórmula utilizada le otorga más peso a los *clusters* de mayor tamaño.

Queda por ver si las relaciones antes establecidas respecto de la precisión de la inferencia se mantienen al incorporar al análisis una segunda vía de clasificación, lo que va a ejemplificarse con la variable OCUPADO. En este caso, para realizar inferencia clásica sería necesario conocer la cantidad total de alumnos por facultad de cada género que se encuentran ocupados y desocupados. Como ello no es factible, se propone aplicar las proporciones muestrales por carrera al total de individuos de cada *cluster*

Las medias estimadas por género y situación ocupacional se presentan en la Tabla 14. Al condicionar sobre dos covariables, la inferencia clásica presenta un problema: se pierde información sobre los *clusters* en los cuales m_i tamaño de muestra por *cluster* valga 1 o 0, ya que en tales casos, no puede estimarse la varianza.

Los resultados obtenidos muestran que la inferencia clásica se resiente si se desea estimar la proporción, en este caso de alumnos con vocación emprendedora, contemplando la influencia de más de una covariable. Al incorporar la variable OCUPADO, la inferencia clásica se torna más imprecisa y desperdicia aquellos *clusters* que no brindan suficiente información para

estimar la varianza.

El análisis llevado a cabo demuestra que la precisión no es el único criterio a tener en cuenta al seleccionar el enfoque a utilizar. Evidentemente, múltiples factores relativos a la configuración elegida influyen sobre la amplitud de los intervalos de confianza. Por lo tanto, no es posible establecer que un método sea mejor que otro en cualquier circunstancia.

CONCLUSIONES

A fin de concluir, en términos generales, acerca de la conveniencia de los métodos discutidos, es importante puntualizar las ventajas y desventajas que posee cada uno. El principal aspecto desfavorable de la inferencia basada en modelos es que requiere numerosos recursos de cálculo, al tratarse de métodos computacionalmente intensivos. Siempre que el modelo formulado sea válido, entre las ventajas derivadas de su uso se encuentran que:

- Es posible estimar la media para cualquier combinación de covariables, lo cual otorga gran flexibilidad al análisis.
- La estimación es independiente de los valores poblacionales que suelen ser desconocidos, por lo que no es necesario contar con información adicional a la proveniente de la muestra.
- El modelo utiliza la totalidad de las observaciones para la estimación de unos pocos parámetros.
- La media estimada no depende del tamaño relativo de los *clusters*.

Como contrapartida, realizar inferencia basada en diseño muestral pone de manifiesto las siguientes desventajas respecto de la inferencia basada en modelos:

- Como insumo para estimar la media y la varianza se requiere información acerca del tamaño total del *cluster* para la subpoblación bajo análisis, la cual puede no estar disponible por fallas en los marcos muestrales o por tratarse de covariables que no tienen diseño, cuyos valores se determinan a posteriori de las encuestas realizadas.
- Las varianzas estimadas se calculan con distintos tamaños muestrales según la configuración que se adopte para las covariables, cambiando por consiguiente los niveles de precisión de las estimaciones.
- Se dispone de un tamaño de muestra menor al condicionar sobre una covariable o combinación de covariables, debido a lo cual los intervalos

- de confianza de las proporciones estimadas pueden ser muy amplios.
- Es posible que se pierda información para algunos *clusters*.

Si bien la inferencia clásica ofrece la ventaja de no requerir procedimientos iterativos de estimación, su aplicación conlleva una serie de dificultades desde el punto de vista práctico que la tornan menos atractiva que el uso de modelos:

- Cuando las estadísticas disponibles son deficientes, sea que contengan errores o datos faltantes, es imposible que se muestree exactamente la población hacia la cual se desea inferir. La falta de marcos muestrales completos que coincidan con la población objetivo plantea divergencias entre lo que la investigación se propone y los resultados que efectivamente se alcanzan.
- Si el diseño muestral es complejo, determinar cuál es el estimador más adecuado de la varianza de la proporción impone dificultades adicionales al análisis.

Debe notarse que, bajo ambos métodos, existen problemas inherentes al cumplimiento de los supuestos o relacionados con la naturaleza asintótica de la inferencia. Con la inferencia basada en modelos se depende del cumplimiento de supuestos difíciles o imposibles de verificar y, al ser los métodos asintóticos, se generan dudas acerca de los valores *p* observados. Con la inferencia clásica, la inferencia también es asintótica y las fórmulas a emplear establecen supuestos que no siempre se cumplen por ejemplo, que se conoce el tamaño total del *cluster*.

Sin embargo, en tanto se verifique la validez del modelo, la inferencia basada en modelos otorga una mayor flexibilidad al análisis, con la ventaja de que sus estimaciones se alimentan exclusivamente de la información muestral y la inferencia puede resultar tanto o más precisa que la efectuada con el método clásico.

Privilegiar el uso de la inferencia basada en modelos reviste gran importancia en la práctica, si se tienen en cuenta las desventajas implícitas en la aplicación de la inferencia clásica. Asimismo, se manifiesta como una alternativa que puede disminuir el costo de una investigación ante la falta de buenos marcos de información, al hacer posible optimizar el trabajo de campo sobre la base del conocimiento previo que se tiene de las unidades muestrales.

BIBLIOGRAFÍA

Agresti, A. (2002), *Categorical data analysis*, 2nd ed., John Wiley.

Diggle, P. (ed.) *et al.*, (2002), *Analysis of longitudinal data*, 2nd ed., New York, Oxford University Press.

Fahrmeir, L. & Tutz, G. (2001), *Multivariate statistical modelling based on generalized linear models*. 2nd ed. New York: Springer-Verlag.

Liang, K. & Zeger, S. (1986), *"Longitudinal data analysis using generalized linear models". Biometrika*, 73 (1): 1322.

Lipsitz, S. *et al.* (1994), *"Performance of generalized estimating equations in practical situations". Biometrics*, 50: 270278.

Longford, N. (1994), *"Logistic regression with random coefficients". Computational Statistics and Data Analysis*, 17: 115.

McCullagh, P. & Nelder, J., (1989), *Generalized linear models*, 2nd ed., New York, Chapman & Hall.

Mc Culloch, C. & Searle, S., (2001) *Generalized, linear and mixed models*, John Wiley.

Rodríguez, G. & Goldman, N. (1995), *"An assessment of estimation procedures for multilevel models with binary response". Journal of the Royal Statistical Society*, Ser. A, 158: 7389.

SAS Institute Inc., (1999), *SAS OnlineDoc* [en cd-rom], version 8, Cary, NC, SAS Institute Inc.

Scheaffer, R.; Mendenhall, W. & Ott, L. (1987), *Elementos de muestreo*. México: Grupo Editorial Iberoamérica.

Spiess, M. & Hamerle, A. (2000), *"A comparison of different methods for the estimation of regression models with correlated binary responses". Computational Statistics & Data Analysis*, 33 (4): 439455.

Zeger, S. & Liang, K. (1986), *"Longitudinal data analysis for discrete and continuous outcomes". Biometrics*, 42: 121130.

Zeger, S.; Liang, K. & Albert, P., (1988), *"Models for longitudinal data: a generalized estimating equation approach", Biometrics*, 44, pp. 10491060.

ANEXO

Tabla 1: CARACTERÍSTERICAS DE LAS FACULTADES INCLUIDAS EN LA MUESTRA

Facultad	Carrera	Gestión	Localización
U1	Económicas	Pública	Zona 1
U2	Económicas	Pública	Zona 1
U3	Económicas	Privada	Zona 1
U4	Económicas	Privada	Zona 1
U5	Económicas	Pública	Zona 2
U6	Económicas	Pública	Zona 2
U7	Económicas	Privada	Zona 2
U8	Económicas	Privada	Zona 2
U9	Ingeniería	Pública	Zona 1
U10	Ingeniería	Pública	Zona 1
U11	Ingeniería	Privada	Zona 1
U12	Ingeniería	Privada	Zona 1
U13	Ingeniería	Pública	Zona 2
U14	Ingeniería	Pública	Zona 2

Tabla 2: PROPORCIÓN ESTIMADA DE ALUMNOS CON VOCACIÓN EMPRENDEDORA MEDIANTE INFERENCIA CLÁSICA

			Intervalo de confianza			
			Lím. inferior	Lím. superior		
0.399	0.0007	0.026	0.348	0.452	0.104	723

Tabla 3: PROPORCIONES ESTIMADA DE ALUMNOS CON VOCACIÓN SEGÚN LA GESTIÓN PÚBLICA O PRIVADA DE LA FACULTAD MEDIANTE INFERENCIA CLÁSICA

				Intervalo de confianza			
				Lím. inferior	Lím. superior		
Pública	0.374	0.001	0.035	0.305	0.443	0.138	624
Privada	0.674	0.004	0.065	0.546	0.802	0.256	99

Tabla 4: Proporciones estimadas de alumnos con vocación emprendedora según carrera mediante inferencia clásica

				Intervalo de confianza			
				Lím. inferior	Lím. superior		
Economía y Adm.	0.405	0.001	0.030	0.346	0.464	0.118	402
Ingeniería	0.373	0.002	0.040	0.295	0.452	0.157	321

Tabla 5: Proporciones estimadas de alumnos con vocación emprendedora por género mediante inferencia clásica

				Intervalo de confianza			
				Lím. inferior	Lím. superior		
Mujeres	0.316	0.001	0.033	0.251	0.381	0.130	248
Hombres	0.469	0.001	0.035	0.401	0.538	0.137	475

Tabla 6: Definición de covariables

Covariable	Descripción y codificación	Hipótesis de trabajo
GENERO	Indica si el alumno es hombre (1) o mujer (0).	Los hombres tienen mayores chances de poseer vocación emprendedora
OCUPADO	Indica si el alumno está actualmente trabajando (1) o no (0).	Los individuos ocupados tienen mayores chances de poseer vocación emprendedora.
ACTITUD	Indica si ante una situación de desempleo en el corto plazo el alumno buscaría una idea de negocios (1) u optaría por un trabajo para el cual estuviese sobrecalificado, no vinculado con su profesión o permanecería desempleado (0).	Los alumnos con actitud empresarial frente al desempleo tienen mayores chances de poseer vocación emprendedora.
VISIÓN	Indica si el alumno visualiza la actividad empresarial en forma favorable (1) o desfavorable (0).	Los alumnos con una visión favorable de la actividad empresarial tienen mayores chances de poseer vocación emprendedora.
RIESGO	Indica si el alumno tiene una propensión al riesgo alta (1) o media/baja (0).	Los alumnos con alta propensión al riesgo tienen mayores chances de poseer vocación emprendedora.
CREATIV	Indica si el alumno desarrolla en su tiempo libre alguna actividad creativa (1) creativa alta - o no (0) - creatividad media/baja-	Los alumnos con un alto nivel de creatividad tienen mayores chances de poseer vocación emprendedora

Tabla 7: ESTIMACIÓN DEL MODELO MARGINAL CON ESTRUCTURA DE CORRELACIÓN INTERCAMBIABLE

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* prueba de cuasiscore*
INTERCEPTO	-3.804	0.363	<0.001	0.02	
GENERO	0.930	0.152	<0.001	2.54	0.026
OCUPADO	1.003	0.235	<0.001	2.73	0.028
ACTITUD	1.040	0.130	<0.001	2.83	0.007
VISION	1.503	0.288	<0.001	4.49	0.009
RIESGO	0.766	0.168	<0.001	2.15	0.007
CREATIV	0.520	0.220	0.018	1.68	0.080

Tabla 8: ESTIMACIÓN DEL MODELO MIXTO CON VEROSIMILTUD COMPLETA

Covariable	β	Error Estándar	Valor *p* Test de Wald	exp(β)
INTERCEPTO	-3.937	0.409	<0.001	0.02
GENERO	0.985	0.230	0.001	2.68
OCUPADO	1.035	0.193	<0.001	2.81
ACTITUD	1.085	0.186	<0.001	2.96
VISION	1.598	0.291	<0.001	4.94
RIESGO	0.787	0.197	0.002	2.20
CREATIV	0.547	0.208	0.021	1.73
SIGMA	0.558	0.211	0.020	

Tabla 9: INTERPRETACIÓN DE LOS COEFICIENTES ESTIMADOS BAJO EL MODELO MARGINAL

Covariable	exp(β)	Interpretación
GENERO	2.54	Controlando por las demás covariables, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 2.5 si es de sexo masculino.
OCUPADO	2.73	Controlando por las demás covariables, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 2.7 si está ocupado.
ACTITUD	2.83	Controlando por las demás covariables, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 2.8 si tiene actitud emprendedora frente al desempleo.
VISION	4.49	Controlando por las demás covariables, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 4.5 si visualiza favorablemente la actitud emprendedora.
RIESGO	2.15	Controlando por las demás covariables, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 2.1 si tiene una alta propensión al riesgo.
CREATIV	1.68	Controlando por las demás covariables, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 1.7 si tiene una alta creatividad.

Tabla 10: Interpretación de los coeficientes estimados bajo el modelo mixto

Covariable	exp(β)	Interpretación
GENERO	2.68	Controlando por las demás covariables y para una facultad determinada, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 2.7 si es de sexo masculino.
OCUPADO	2.81	Controlando por las demás covariables y para una facultad determinada, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 2.8 si está ocupado.
ACTITUD	2.96	Controlando por las demás covariables y para una facultad determinada, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 3 si tiene actitud emprendedora frente al desempleo.
VISIÓN	4.94	Controlando por las demás covariables y para una facultad determinada, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 4.9 si visualiza favorablemente la actividad emprendedora.
RIESGO	2.20	Controlando por las demás covariables y para una facultad determinada, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 2.2 si su propensión al riesgo es alta.
CREATIVIDAD	1.73	Controlando por las demás covariables y para una facultad determinada, las chances de que un alumno posea vocación emprendedora se multiplican por un factor de 1.7 si tiene alta creatividad.

Tabla 11: Probabilidades estimadas y riesgos relativos bajo el modelo marginal con estructura de coreelación intercambiable

Nivel covariables	Pr(VE=1)	Riesgo relativo
Categoría modal	0.409	
Si el alumno es mujer	0.21	0.52
Si el alumno no está trabajando	0.20	0.50
Si el alumno posee actitud empresarial frente al desempleo	0.66	1.62
Si el alumno posee una visión desfavorable de la actividad empresarial	0.13	0.33
Si el alumno es propenso al riesgo	0.60	1.46
Si el alumno posee alta creatividad	0.54	1.32

Tabla 12: EFECTO DE LA FACULTAD SOBRE LA PROBABILIDAD DE PRESENCIA DE VOCACIÓN EMPRENDEDORA BAJO EL MODELO MIXTO CON VEROSIMILITUD COMPLETA

Categoría modal	$Pr(VE=1/U_i)$	Riesgo relativo
Efecto aleatorio nulo	0.421	
U1	0.49	1.18
U2	0.35	0.82
U3	0.63	1.49
U4	0.57	1.36
U5	0.30	0.72
U6	0.38	0.90
U7	0.51	1.21
U8	0.45	1.07
U9	0.25	0.60
U10	0.38	0.90
U11	0.42	1.00
U12	0.52	1.23
U13	0.37	0.88
U14	0.33	0.78

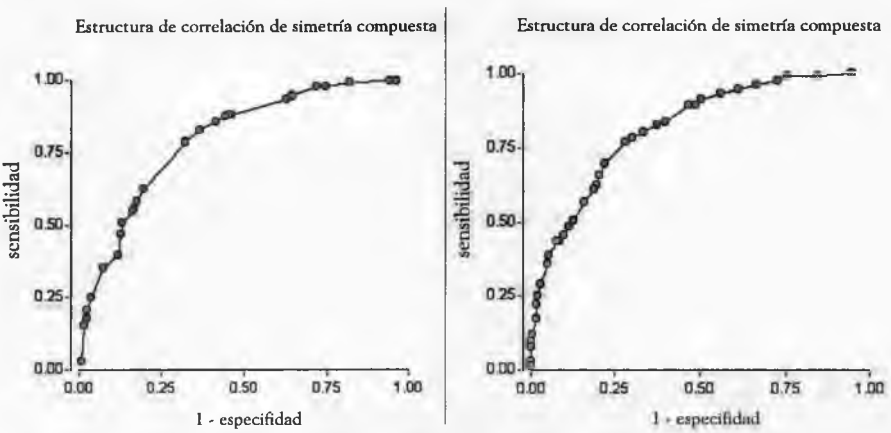

Figura 1: CURVAS ROC CORRESPONDIENTES A LOS MODELOS MARGINAL Y MIXTO

Tabla 13: PROPORCIONES ESTIMADAS DE ALUMNOS CON VOCACIÓN EMPRENDEDORA POR GÉNERO BAJO AMBOS MÉTODOS DE INFERENCIA

Subpoblación	Inferencia	Proporción estimada	Amplitud del intervalo de confianza
	Inferencia Clásica	0.469	0.137
	Enfoque marginal Corr. intercambiable	0.414	0.143
	Enfoque mixto $U_i=0$	0.424	0.226
	Inferencia Clásica	0.316	0.130
	Enfoque marginal Corr. intercambiable	0.188	0.107
	Enfoque mixto $U_i=0$	0.184	0.169

Tabla 14: PROPORCIONES ESTIMADAS DE ALUMNOS CON VOCACIÓN EMPRENDEDORA POR GÉNERO Y SITUACIÓN OCUPACIONAL BAJO AMBOS MÉTODOS DE INFERENCIA

Subpoblación	Inferencia	Proporción estimada	Amplitud del intervalo de
	Inferencia clásica	0.590	0.204
	Enfoque marginal Corr. intercambiable	0.535	0.208
	Enfoque mixto $U_i=0$	0.550	0.240
	Inferencia clásica	0.399	0.228
	Enfoque marginal Corr. intercambiable	0.257	0.164
	Enfoque mixto $U_i=0$	0.256	0.214
	Inferencia clásica	0.341	0.132
	Enfoque marginal Corr. intercambiable	0.297	0.132
	Enfoque mixto $U_i=0$	0.303	0.220
	Inferencia clásica	0.137	0.121
	Enfoque marginal Corr. intercambiable	0.112	0.079
	Enfoque mixto $U_i=0$	0.097	0.112