UNIVERSIDAD NACIONAL DE CÓRDOBA

Título de la tesis:

Análisis de encuestas basado en diseño y modelos muestrales: una comparación entre métodos de inferencia aplicados al estudio de la vocación emprendedora en alumnos universitarios

Para optar al grado de:

Magister en Estadística Aplicada

Autora:

Lic. Natacha Liseras

2004

COMISIÓN ASESORA DE TESIS

Director:

Dr. Raúl Macchiavelli
UPR – Universidad de Puerto Rico

Miembros:

Dra. Mónica Balzarini
UNC – Universidad Nacional de Córdoba

Dr. José Vila Gisbert
UV – Universidad de Valencia

Defensa de tesis:

15 de abril de 2004

Agradecimientos

En primer lugar, quiero agradecerle al Dr. Raúl Macchiavelli su confianza, su entusiasmo y su maravillosa dedicación. A la Dra. Mónica Balzarini le agradezco su participación y sus valiosos aportes. Para ambos mi gratitud por todo lo que me han enseñado y por su calidez que hizo muy gustosas las horas de trabajo compartidas.

Por su tiempo, sugerencias y bibliografía brindada, le doy las gracias a la Dra. María del Pilar Díaz, a la Dra. Margarita Díaz, al MSc. Julio di Rienzo, al Dr. Fernando Ferrero, al Dr. Raúl Martínez y al Dr. Walter Robledo. En particular, al Dr. José Vila Gisbert le agradezco haber soportado mis insistentes preguntas sobre la fórmula de la varianza, su grata colaboración y sus acertados comentarios.

En lo personal, quiero expresar mi agradecimiento a Ana Gennero de Rearte por haberme dado la oportunidad de cursar esta maestría y por estimularme a seguir estudiando; a mis amigas y a mis padres por todo su apoyo; a Miriam Berges por su incondicionalidad; a Cecilia Bruno por comportarse como una amiga; a los integrantes del Centro de Documentación de mi facultad por las mil búsquedas bibliográficas realizadas. Y, muy especialmente, gracias a Leo por acompañarme en todos mis proyectos y alentarme a hacer lo que me gusta.

Resumen

En el presente trabajo se comparan los métodos de análisis de encuestas propios de la inferencia clásica y de la 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 las funciones de verosimilitud completa y condicional, con enlace logístico.

El diagnóstico del modelo no resulta posible debido a: *(a)* la falta de una función de verosimilitud bajo el enfoque marginal, lo que inhibe el uso de las medidas basadas en el cociente de verosimilitud; *(b)* la naturaleza binaria de la variable respuesta y de las covariables, que torna poco informativo el uso de las técnicas usualmente empleadas en modelación. Entre otras alternativas, se propone el cálculo de una tasa de error por validación cruzada *leave-one-out*, a fines de evaluar la predictibilidad de los modelos estimados.

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, estimándose 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.

Palabras claves

Observaciones binarias correlacionadas – Inferencia basada en modelos – Ecuaciones de estimación generalizadas – Modelos marginales – Modelos de efectos aleatorios – Vocación emprendedora

Analyzing surveys by using design-based and model-based methods:

A comparison of inferential methods in a study of entrepreneurship

Abstract

In this thesis we compare the design-based and model-based inferential methods for survey analysis of correlated binary data. We propose the formulation of marginal generalized linear models and mixed generalized linear models with complete and conditional likelihood functions, using logit link.

It is not possible to diagnose the model because: *(a)* the lack of a likelihood function under the marginal approach, which makes invalid the use of measures based on the likelihood ratio; *(b)* the binary nature of both the response and explanatory variables, so the use of *deviance*, residuals and graphic techniques are uninformative. Among other alternatives, we propose to calculate a leave-one-out cross-validation error rate to evaluate the predictive power of the estimated models.

The application consists of a cross-sectional study in which dependence between observations arises because of the subsampling of primary units –students surveyed within universities–. The response variable is the presence of entrepreneurial vocation among undergraduates students of Economics, Business Administration and Engineering. The estimated proportion of students with entrepreneurial vocation is 0.4 under both inferential methods, being the model-based estimation more flexible than the design-based one.

Keywords

Correlated binary data – Model-based inference – Generalized estimating equations – Marginal models – Random-effects models – Entrepreneurial vocation

TABLA DE CONTENIDOS

1. INTRODUCCIÓN

En las Ciencias Sociales, 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 aún 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 a través de 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, e.g. 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[1]. Su uso representa un importante aporte metodológico para las Ciencias Sociales, dado que su campo usual de aplicación es el de las Ciencias Biológicas.

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)*; 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. Si se formula un modelo mixto con verosimilitud condicional, se supone a los efectos aleatorios como parámetros auxiliares *(nuisance)* y se los condiciona fuera del modelo.

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

[1] La distinción entre estos enfoques es irrelevante si la variable respuesta tiene distribución normal, pero con variables binarias la combinación de modelos logísticos específicos para cada *cluster* –o grupo de observaciones no independientes– no es, en general, un modelo logístico para la población (Fahrmeir & Tutz, 2001).

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.

A continuación se plantean los objetivos (sección 2) y se desarrolla el marco conceptual (sección 3), en donde se discuten los métodos propios de la inferencia clásica y basada en modelos, haciendo hincapié en los modelos para datos correlacionados. Luego se describen la metodología de trabajo (sección 4), la aplicación y el diseño muestral llevado a cabo (sección 5). Posteriormente, se aplican la inferencia clásica (sección 6) y la inferencia basada en modelos (sección 7) al caso de estudio. Seguidamente, se comparan los resultados hallados por ambos métodos (sección 8), resaltando las ventajas y desventajas de cada una de estas estrategias de análisis. Por último, se presentan las conclusiones (sección 9) y se plantean los aspectos que quedan pendientes para futuras investigaciones (sección 10).

2. OBJETIVOS

Los principales objetivos de esta tesis 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.

DISCUSIÓN TEÓRICA DE LOS MÉTODOS DE INFERENCIA

- Inferencia basada en diseño muestral
- Inferencia basada en modelos
  - Modelos lineales generalizados
  - Modelación de observaciones binarias correlacionadas
  - Enfoque marginal
  - Enfoque mixto
  - Relación entre los enfoques marginal y mixto
  - Inferencia y diagnóstico del modelo

3. MARCO CONCEPTUAL

3.1. Inferencia basada en diseño muestral

El diseño muestral desempeña un rol vital en toda investigación. La inferencia basada en diseño muestral involucra dos procesos: el de selección y el de estimación. Mientras que en el primero se determina qué individuos serán incluidos en la muestra, en el segundo se calculan los estadísticos muestrales que estiman los valores poblacionales (Kish, 1965).

Las propiedades del muestreo clásico se definen en términos de su comportamiento en muestreos repetidos y descansan en lo que se denomina el principio de representatividad. Éste establece que cada unidad contenida en la muestra se representa a sí misma y a un grupo de unidades no muestreadas, cuyas propiedades se hallan próximas. En términos generales, para cualquier muestra aleatoria seleccionada con reemplazo, la ponderación asignada a cada unidad presente en la misma debe ser igual al recíproco de su probabilidad de inclusión (Brewer, 1999).

La cuestión fundamental consiste en inferir a partir de la muestra hacia una población de mayor tamaño. Por consiguiente, es preciso contar con un marco de información cuya diferencia entre la población objetivo y la población muestreada sea lo suficientemente pequeña, ya que las conclusiones extraídas de la muestra sólo son aplicables a esta última (Cochran, 1980).

Cada diseño muestral provee estimaciones con distinto error estándar, por lo cual uno de los principales objetivos es elegir el diseño que minimice dicho error. Un buen diseño contempla numerosos aspectos, tales como (Kish, 1965):

- La definición de las variables, debiendo especificarse la naturaleza de las características a medir, las reglas de clasificación y las unidades en las que serán expresadas.
- Los métodos de observación, incluyendo la recolección y el procesamiento de los datos.

- Los métodos de análisis que transforman los datos relevados en resultados comprensibles y utilizables.
- La utilización de los resultados, que puede traducirse en decisiones concretas o pasar a formar parte del conocimiento.
- La precisión deseada de los resultados, la cual viene determinada por decisiones estadísticas relacionadas al diseño muestral.

Una fase importante del diseño es la elección del tamaño de la muestra, decisión que sólo puede tomarse satisfactoriamente si se dispone de información suficiente para juzgar cuál es el tamaño óptimo (Cochran, 1980). La fórmula aplicable para definir el tamaño de la muestra para una proporción, bajo un muestreo aleatorio simple sin reemplazo, es:

$$n \geq \frac{z^2 \ \hat{\mu} \ (1 \ - \ \hat{\mu})}{e^2} \left(\frac{N-n}{N-1} \right) , \qquad [1]$$

siendo:

- n el tamaño de la muestra.
- N el tamaño de la población.
- e el error de estimación.
- z el cuantil $(1-\alpha)$ de la distribución normal estándar.
- $\hat{\mu}\left(1-\hat{\mu}\right)$ la varianza binomial.
- $\left(\frac{N-n}{N-1} \right)$ el factor de corrección para poblaciones finitas, si se considera que la relación (n / N) no es despreciable.

Como surge de [1], el cálculo del tamaño muestral no depende del tamaño de la población –excepto a través de la fracción de muestreo–, pero sí del valor de la proporción $\hat{\mu}$ que desea estimarse. Esto se explica porque con variables Bernoulli la media o proporción no es independiente de su varianza. Si se desea estimar la proporción para distintas subdivisiones o dominios dentro de la población, el cálculo del tamaño muestral debe hacerse por separado para cada subdivisión y el tamaño total surge por adición.

También debe elegirse una probabilidad $(1 - \alpha)$ que especifique la fracción de veces en

muestreos repetidos en que el error de estimación será menor al establecido. Debido a que los estimadores para muestras grandes tienen distribución aproximadamente normal por el Teorema Central del Límite, es que puede utilizarse el cuantil α de la distribución normal para establecer la amplitud del intervalo de confianza para la proporción.

Por último, es necesario definir un límite para el error de estimación, el cual equivale a la máxima diferencia en valor absoluto entre el estimador y el parámetro:

$$e = \left| \hat{\mu} - \mu \right| . \qquad [2]$$

La elección del nivel de error no es una decisión estadística, sino que depende del problema en estudio. Aún cuando existe un error muestral y otro no muestral, el que se tiene en cuenta en esta instancia es el primero, ya que el segundo no puede controlarse incrementando el tamaño de la muestra (Cornfield, 1951).

En el muestreo sin reemplazo, la proporción cambia a lo largo del proceso de obtención de unidades. Sin embargo, si se considera que la proporción permanece constante, el proceso para conformar la muestra consiste en una serie de ensayos con igual probabilidad que determina una distribución de frecuencias binomial. Si la variable respuesta *Y* es de naturaleza binaria –i.e., asume el valor 1 cuando la característica buscada se halla presente en el individuo o el valor 0 si la misma está ausente–, el número de individuos que poseen la característica bajo análisis en la población y en la muestra son, respectivamente:

$$A = \sum_{i=1}^{N} y_i , \qquad [3]$$

$$a = \sum_{i=1}^{n} y_i , \qquad [4]$$

donde y_i denota el estado de *Y* en el i-ésimo individuo. La proporción de individuos con la característica de interés, en la población y en la muestra, están dadas por:

$$\mu = \frac{A}{N} = \frac{\sum_{i=1}^{N} y_i}{N} , \qquad [5]$$

$$\hat{\mu} = \frac{a}{n} = \frac{\sum_{i=1}^{n} y_i}{n}. \qquad [6]$$

La proporción estimada tiene una distribución muestral que se aproxima a la normal en muestras grandes. Bajo las condiciones de un muestreo aleatorio simple, el estimador $\hat{\mu}$ es insesgado, expresándose la varianza como:

$$Var\left(\hat{\mu}\right) = \frac{\hat{\mu}\left(1-\hat{\mu}\right)}{\left(n-1\right)}. \qquad [7]$$

Si el tamaño de la muestra representa más del 5% de la población, la expresión anterior debe multiplicarse por el factor de corrección para poblaciones finitas:

$$Var\left(\hat{\mu}\right) = \frac{\hat{\mu}\left(1-\hat{\mu}\right)}{\left(n-1\right)}\left(\frac{N-n}{N-1}\right). \qquad [8]$$

Cuando el muestreo es por conglomerados *(clusters)* en una etapa, la potencia del diseño depende más del número de conglomerados que del tamaño de la muestra. La estimación de la media de cada *cluster* surge de dividir el total de casos que presentan la característica buscada sobre el tamaño del conglomerado:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}, \qquad [9]$$

donde:

- El subíndice *i* representa al *cluster (i = 1, ..., k)*.
- El subíndice *j* representa al individuo *(i = 1, ..., m_i)*.
- m_i indica el tamaño del i-ésimo *cluster*.

La media global, que estima la proporción poblacional, es ligeramente sesgada:

$$\hat{\mu}_{\cdot} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m_i} y_{ij}}{\sum_{i=1}^{k} m_i} . \qquad [10]$$

Una posible expresión para la varianza del estimador bajo un muestreo por conglomerados es (Cochran, 1980):

$$Var\left(\hat{\mu}_{\cdot}\right) = \frac{(1-f)}{k\overline{m}^2} \cdot \frac{\sum_{i=1}^{k} m_i^2 \left(\hat{\mu}_i - \hat{\mu}_{\cdot}\right)^2}{k-1} , \qquad [11]$$

siendo:

- $\overline{m}$ el tamaño promedio del *cluster*.
- k el número de *clusters* en la muestra
- $f = \frac{k}{K}$ la fracción de muestreo, con K el número de *clusters* en la población.

Si los individuos pertenecientes a los conglomerados 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. En tal caso, debe disponerse de marcos que listen todas las unidades de selección primaria que hay en la población, así como las unidades en cada uno de los conglomerados seleccionados.

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 (ver Anexo B).

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

$$\hat{\mu}_{..} = \frac{\sum_{i=1}^{k} M_i \hat{\mu}_{i.}}{\sum_{i=1}^{k} M_i}. \qquad [12]$$

Como M_i representa el número de elementos en el i-ésimo *cluster*, el estimador descripto en [*12*] 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 $\overline{M}^2$ se calcula con información proveniente de la muestra. La varianza estimada de la media global es:

$$Var\left(\hat{\mu}_{..}\right) = \left(\frac{K-k}{K}\right)\left(\frac{1}{k\overline{M}^2}\right)S_1^2 + \frac{1}{kK\overline{M}^2}\sum_{i=1}^{k} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\left(\frac{\hat{\mu}_{i.}\left(1-\hat{\mu}_{i.}\right)}{m_i - 1}\right) =$$
$$= \left(1-f_1\right)\left(\frac{1}{k\overline{M}^2}\right)S_1^2 + \frac{1}{kK\overline{M}^2}\sum_{i=1}^{k} M_i^2\left(1-f_{2i}\right)\left(\frac{\hat{\mu}_{i.}\left(1-\hat{\mu}_{i.}\right)}{m_i - 1}\right), \qquad [13]$$

$$S_1^2 = \frac{\sum_{i=1}^{k} M_i^2\left(\hat{\mu}_{i.} - \hat{\mu}_{..}\right)^2}{k-1}, \qquad [14]$$

siendo:

- $\overline{M}$ el tamaño promedio del *cluster*.
- k el número de *clusters* en la muestra.
- K el número de *clusters* en la población.
- f_1 la fracción de muestreo de primera etapa.
- f_{2i} la fracción de muestreo de segunda etapa, distinta para cada *cluster*.
- 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:

$$z = \frac{\hat{\mu}_{..}^{(1)} - \hat{\mu}_{..}^{(2)}}{\sqrt{Var\left(\hat{\mu}_{..}^{(1)}\right) + Var\left(\hat{\mu}_{..}^{(2)}\right)}} \sim N(0,1). \qquad [15]$$

3.2. Inferencia basada en modelos

Una alternativa a la inferencia clásica –basada en el diseño muestral– para la estimación de parámetros poblacionales, es la inferencia basada en modelos *(model-based inference)*. Su estudio fue iniciado por Royall (1970), quien argumenta que el análisis basado en el diseño de muestreo utiliza una estructura de probabilidades incorrecta, no relacionada con los datos mismos sino con la manera en que éstos son recogidos[2]. Sobre la base del principio de condicionalidad –que Birnbaum (1962) define como "la irrelevancia de experimentos no realizados efectivamente"–, este enfoque establece que para inferir sólo es relevante la muestra observada y no el conjunto de posibles muestras que pueden seleccionarse con un diseño determinado[3].

Este tipo de inferencia debe su nombre a que la información recolectada en la muestra se utiliza para construir modelos, y el análisis se efectúa en relación a los parámetros del mismo. Optar entre un modelo de efectos fijos o un modelo mixto depende de aspectos vinculados con la selección muestral. Si la variabilidad en las respuestas es consistente con el supuesto de individuos seleccionados al azar de una población de gran tamaño, para hacer inferencia deben utilizarse modelos que permitan calificar de aleatoria dicha variabilidad (Beitler & Landis, 1985).

Bajo esta estrategia de análisis, la inferencia puede efectuarse a partir de los modelos lineales clásicos o de los generalizados. En el caso de respuestas binarias correlacionadas, ello requiere el uso de extensiones a los modelos lineales generalizados que serán comentadas en los apartados 3.2.3 y 3.2.4.

Si existe interés en estimar la proporción de individuos que presentan una característica determinada para los distintos niveles de una variable de clasificación, bajo la inferencia clásica estas proporciones sólo pueden calcularse particionando la muestra total. En cambio, la inferencia basada en modelos ofrece la ventaja de estimar los parámetros de interés con la totalidad de las observaciones de la muestra. Por consiguiente, cabe esperar que esto redunde en una mayor precisión en el proceso de estimación.

[2] Royal (1970) es citado por Brewer (1999).

3.2.1. Modelos lineales generalizados

Durante décadas, los modelos lineales del tipo:

$$Y = X\beta + e \,, \qquad [16]$$

donde:

- Y es una variable respuesta.
- β es un vector de parámetros asociados a los efectos que explican variabilidad en la respuesta.
- X una matriz de diseño que asocia a la variable Y con β.
- e un término de error aleatorio.

han conformado la base de la mayoría de los análisis para datos con distribución normal. Métodos análogos a los desarrollados para los modelos lineales pueden aplicarse cuando la variable respuesta tiene una distribución distinta a la normal o cuando la relación entre ella y las covariables, o efectos, no es de la forma lineal indicada en [*16*] (Dobson, 1983).

Muchas de las propiedades deseables de la distribución normal son compartidas por una clase más amplia de distribuciones que pertenecen a la familia exponencial uniparamétrica. La familia exponencial representa una estructura matemática general que incluye como casos especiales a distintas distribuciones de probabilidad discretas y continuas –binomial, Poisson, normal, normal inversa y gama–. Dada una variable aleatoria Y cuya función de probabilidad o densidad depende de un único parámetro de interés θ, se dice que ésta pertenece a la familia exponencial si puede expresarse de la forma:

$$f(y_{ij} \;/\; \phi) \;=\; \exp\left[\sum_{i=1}^{k}\sum_{j=1}^{m} h(y_{ij})\;\theta_{ij} \;-\; \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{b_{ij}(\theta_{ij})}{a(\phi)} \;+\; c_i\;(y_{ij},\;\phi)\right], \qquad [17]$$

siendo:

- θ el parámetro natural de la distribución.
- ϕ un parámetro de escala o de sobredispersión que se supone conocido.

[3] Birnbaum (1962) es citado por Brewer (1999).

- $a(\phi) = \frac{\phi}{w_{ij}}$, donde w_{ij} representa pesos "a priori" y $a(\cdot)$ se considera conocida.

Si $h(y_{ij}) = y_{ij}$ se dice que la distribución tiene la forma canónica. Si incluye otros parámetros además de θ, éstos se consideran como ruido *(nuisance)* y se tratan como si fueran conocidos. A partir de las propiedades de la familia exponencial, la función generadora de momentos y la de acumulantes, puede demostrarse que (McCullagh & Nelder, 1989):

- La esperanza de la variable respuesta equivale a la primera derivada de $b(\theta_{ij})$:

$$\mu_{ij} = E\left(y_{ij}\right) = b'\left(\theta_{ij}\right). \qquad [18]$$

- La función de varianza equivale a la segunda derivada de $b(\theta_{ij})$:

$$V(\mu_{ij}) = b''(\theta_{ij}) = \frac{\partial \mu_{ij}}{\partial \theta_{ij}}. \qquad [19]$$

- La varianza de la variable respuesta equivale a una función de la media –a través de la función de varianza– y del parámetro de escala:

$$Var\left(y_{ij}\right) = v_{ij} = a(\phi)\, b''\left(\theta_{ij}\right) = a(\phi)\, V\left(\mu_{ij}\right). \qquad [20]$$

La función de varianza, que se utiliza para describir la variabilidad no sistemática, representa la dependencia de la varianza de la variable respuesta respecto de los parámetros de posición (media) y de escala (varianza). Su relación con $b\left(\theta_{ij}\right)$ establece explícitamente que la función de varianza es condicional a la media (Gill, 2001).

Dado un conjunto de variables aleatorias independientes $\{y_1, ..., y_n\}$ perteneciente a la familia exponencial, los parámetros θ_{ij} carecen de interés directo en la especificación del modelo. Para un modelo lineal generalizado (MLG), se considera un conjunto de p parámetros $\boldsymbol{\beta} = \{\beta_1, ..., \beta_p\}$ con $p < N$, tales que una combinación lineal de los mismos es igual a una función del valor esperado de la variable respuesta:

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^{'}\beta, \qquad [21]$$

siendo:

- g una función conocida, monótona y diferenciable llamada enlace.

- η el predictor lineal.

De este modo, los MLGs pueden definirse como modelos lineales para la media transformada de una variable respuesta que posee una distribución en la familia exponencial. Éstos quedan definidos por[4]:

- Una **componente aleatoria**, dada por las variables aleatorias independientes y_{ij} que poseen una misma distribución perteneciente a la familia exponencial.
- Una componente sistemática o **predictor lineal** en los parámetros que postula la relación entre las covariables.
- Una **función de enlace** que liga a la media de la variable respuesta con las covariables.

Es sólo a través de la función de enlace aplicada al predictor lineal que las covariables ejercen efecto sobre la variable respuesta. El enlace canónico es aquél para el cual el parámetro canónico iguala al predictor lineal:

$$\theta_{ij} = g\left(\mu_{ij}\right) = \eta_{ij} . \qquad [22]$$

Aún cuando los enlaces canónicos poseen propiedades deseables, particularmente en muestras pequeñas, McCullagh & Nelder (1989) expresan que no hay ninguna razón a priori por la cual los efectos sistemáticos de un modelo deban ser aditivos en la escala dada por dicho enlace. Asimismo, estos autores plantean que la elección de la escala es un aspecto importante en la selección de un modelo.

La escala adecuada para el análisis clásico de regresión lineal, e.g., es aquélla que combina varianza constante, errores independientes distribuidos aproximadamente normales y aditividad en los efectos sistemáticos. Sin embargo, con la introducción de los MLGs la normalidad y la varianza constante ya no son necesarias y la aditividad de los efectos, aunque importante, puede lograrse en otra escala. Sí debe conocerse la manera en la que varianza depende de la media (relación media-varianza), mientras que la aditividad se postula como propiedad de una cierta función de la esperanza de la variable respuesta.

En un MLG, la variabilidad residual corresponde a la distribución muestral –descripta mediante la función de varianza– y a la existencia de sobredispersión. Esta última puede modelarse (Palmgren & Ripatti, 2002):

- Con el parámetro ϕ, cuando no hay información disponible respecto de su origen y este único parámetro captura la variabilidad adicional.
- Seleccionando otra distribución de probabilidad.
- Introduciendo un efecto aleatorio en el predictor lineal.

Dado que la función de enlace permite que la relación entre la media de la variable respuesta y el predictor sea no lineal, se introduce una dificultad adicional en el algoritmo de estimación. Con observaciones independientes, el método de máxima-verosimilitud (MV) –que maximiza la verosimilitud o la función de log-verosimilitud (función soporte) de los parámetros para las observaciones– provee estimadores consistentes y asintóticamente normales y eficientes. Éstos representan el valor de los parámetros para los cuales la muestra observada tiene la mayor probabilidad de ocurrencia.

Bajo condiciones muy generales, satisfechas por la familia exponencial de distribuciones, los procesos iterativos encuentran el modo de la función de verosimilitud y dan como resultado un vector de estimadores por MV[5]. Asimismo, la matriz estimada de covarianzas converge en probabilidad a la matriz poblacional, tal como se desea (Gill, 2001). Una de las razones por las cuales los MLGs se restringen a la familia exponencial de distribuciones para *Y*, es debido a que el mismo algoritmo se aplica para cualquier elección de la función de enlace (Agresti, 2002).

Regresión logística

Un enlace comúnmente utilizado cuando los errores poseen distribución Bernoulli es el logístico, el cual expresa al *logit* de la media o logaritmo de las chances *(odds)* como una

[4] Gill (2001) menciona que, si bien tradicionalmente se describe a los MLGs mediante estas tres componentes, en realidad existe una cuarta componente dada por los residuos, los cuales son un determinante crítico para juzgar la calidad del modelo.
[5] Tal es el caso del proceso iterativo de Newton-Raphson, el cual utiliza la matriz de valores esperados de las derivadas segundas de la función de log-verosimilitud (Díaz & Demetrio, 1998).

función lineal de los parámetros. Mientras que la probabilidad debe pertenecer al intervalo (0, 1), el *logit* puede ser cualquier número real. Debe tenerse presente que las observaciones binarias resultan de una relación no lineal entre la probabilidad *(μ_{ij})* y las covariables *(X)*. Por lo tanto, un cambio en las *X* tiene menos impacto si la probabilidad está próxima a 0 o 1 que si se encuentra cerca de 0.5 (Agresti, 2002):

$$\text{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right), \qquad [23]$$

donde:

- μ_{ij} es la probabilidad de éxito.
- $(1 - \mu_{ij})$ es la probabilidad de fracaso.

La función de enlace inversa se emplea para describir la relación entre el predictor lineal y la media de la variable respuesta. En este caso particular resulta ser:

$$g^{-1}(\eta_{ij}) = \mu_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})} = \frac{1}{1 + \exp(-\eta_{ij})} \quad . \qquad [24]$$

Si bien con observaciones binarias puede optarse por distintos enlaces –tales como el probit o el complemento log-log[6]–, el logístico suele preferirse por distintos motivos (Agresti, 1996; Pendergast *et al.*, 1996):

- ⊡ Es el enlace canónico, de modo tal que pueden obtenerse estadísticos suficientes y construirse funciones de verosimilitud condicionales.
- ⊡ En tanto que el intercepto sea tratado como un parámetro de ruido, la interpretación de los coeficientes no varía sea el estudio prospectivo o retrospectivo.
- ⊡ La interpretación de los coeficientes de regresión se efectúa en términos de cocientes de chances, lo cual facilita la comprensión de los resultados.

[6] El enlace probit es $\Phi^{-1}(\mu_i) = X'\beta$, siendo Φ la función de distribución normal. El complemento log-log es de la forma $log\{-log(1 - \mu_i)\} = X'\beta$.

3.2.2. Modelación de observaciones binarias correlacionadas

La existencia de observaciones correlacionadas se manifiesta en numerosos campos de estudio y, cualquiera sea el diseño de la experiencia, es recomendable incorporar esta correlación en el modelo (Larsen *et al.*, 2000; Neuhaus *et al.*, 1991). Tal como señalan Zeger *et al.* (1988), la falta de independencia puede pensarse como una fuente de oportunidades y de desafíos adicionales al análisis.

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).

A su vez, si la variable respuesta es de naturaleza binaria, puede haber sobredispersión. Dicho efecto es propio de todo cuerpo de datos cuya estructura sea jerárquica o anidada, hallándose las mediciones individuales organizadas en conglomerados que pueden pertenecer, a su vez, a unidades de mayor tamaño. Múltiples causas pueden explicar su presencia, siendo las principales: *(a)* que haya una heterogeneidad no captada por las covariables y, por lo tanto, no modelada en el predictor lineal; *(b)* que las respuestas dentro del *cluster* estén positivamente correlacionadas. Puesto que ambas causas se traducen en la existencia de correlación positiva, y ésta contribuye a la varianza, se verifica mayor dispersión que si las observaciones fueran independientes (Fahrmeir & Tutz, 2001).

La solución tradicional cuando las observaciones se hallan correlacionadas consiste en aplicar el modelo jerárquico beta-binomial. Considerando que las observaciones de cada *cluster* son variables Bernoulli independientes idénticamente distribuidas (iid), con una probabilidad de éxito que sigue una distribución beta, el modelo beta-binomial asume que (Pendergast *et al.*, 1996):

- Las respuestas dentro de cada *cluster*, condicional a μ_i, son independientes y tienen una probabilidad común μ_i.
- μ_i sigue una distribución beta con media μ y varianza $\delta\mu(1-\mu)$, donde δ representa al parámetro de sobredispersión.

Si no se condiciona, el número total de respuestas positivas dentro de un *cluster* $Y_{i\cdot} = Y_{i1} + ... + Y_{im}$, sigue una distribución beta-binomial con:

$$E(Y_{i\cdot}) = m_i \mu_i , \qquad [25]$$

$$Var(Y_{i\cdot}) = m_i \mu_i (1-\mu_i)\{1+(m_i-1)\delta\} , \qquad [26]$$

indicando δ la correlación entre cada par de respuestas binarias de un mismo conglomerado.

Es posible imponer un modelo paramétrico a las medias de cada *cluster* dentro del marco de la distribución beta-binomial asumiendo, e.g., que μ_i depende de las covariables a través de una función logística (Diggle *et al.*, 2002). Sin embargo, la interpretación de los efectos de las covariables se torna compleja debido a que esta distribución no es un miembro de la familia exponencial. Ello dificulta relacionar la escala de la distribución beta con la escala logística, así como las distintas distribuciones beta –cada una con sus propios parámetros– entre sí (Longford, 1994).

Una alternativa al uso del modelo beta-binomial consiste en formular modelos que incorporen covariables, además de contemplar la dependencia entre las observaciones. Tal es el caso de los modelos lineales generalizados marginales y mixtos, cuyos coeficientes, si se opta por el enlace *logit*, se interpretan en términos de cocientes de chances. Luego de definir el riesgo relativo y los cocientes de chances se caracterizan ambas estrategias.

Riesgo relativo y cociente de chances

Existen tres estadísticos con los cuales se pueden comparar las respuestas binarias de dos grupos: *(a)* la diferencia de proporciones; *(b)* el riesgo relativo; *(c)* el cociente de chances.

La diferencia de proporciones ha sido comentada en el apartado 3.1. El riesgo relativo *(RR)* equivale al cociente entre las probabilidades de éxito en cada grupo:

$$RR = \frac{\mu_1}{\mu_2}, \qquad [27]$$

que puede asumir cualquier valor real no negativo.

En cuanto al cociente de chances, antes de definirlo es necesario referirse al concepto de chances. Si *Y* es una variable respuesta de naturaleza binaria que asume el valor 1 cuando la característica bajo análisis se encuentra presente (éxito) y 0 cuando la misma se halla ausente (fracaso), se denomina chances *(odds)* al cociente entre las probabilidades de éxito y de fracaso. Puesto que las chances son siempre no negativas, un valor superior a la unidad indica que el éxito es más probable que el fracaso:

$$chances = \frac{\mu_i}{1-\mu_i} = \frac{\Pr(Y=1)}{\Pr(Y=0)}. \qquad [28]$$

Cuando se comparan dos grupos de observaciones binarias, el cociente de chances *(odds ratio - OR)* indica la relación entre las chances en ambos grupos. Si las chances son idénticas –lo que implica que las probabilidades de éxito también lo son–, *OR* será igual a 1; si las chances de éxito son mayores en el primer grupo que en el segundo, *OR* será mayor que 1 y viceversa.

Estas medidas de asociación son, asimismo, los parámetros correspondientes a los efectos para datos binarios. En una regresión logística, el logaritmo de las chances se denomina *logit* y, si hay una única covariable dicotómica, resulta que:

$$\text{logit}(\mu_i / X=1) = \log\left(\frac{\Pr(Y=1/X=1)}{\Pr(Y=0/X=1)}\right) = \beta_0 + \beta_1, \qquad [29]$$

$$\text{logit}(\mu_i / X=0) = \log\left(\frac{\Pr(Y=1/X=0)}{\Pr(Y=0/X=0)}\right) = \beta_0. \qquad [30]$$

El cociente de chances, que se define como la relación entre las chances para aquellas observaciones con *X=1* y aquéllas con *X=0*, equivale al cociente de los *logits*, siendo el logaritmo del cociente de chances igual a β_1:

$$\log(OR) = \log\left(\frac{\left(\frac{\Pr(Y=1/X=1)}{\Pr(Y=0/X=1)}\right)}{\left(\frac{\Pr(Y=1/X=0)}{\Pr(Y=0/X=0)}\right)}\right) = \text{logit}(\mu_i / X=1) - \text{logit}(\mu_i / X=0) = \beta_0 + \beta_1 - \beta_0 = \beta_1 . \quad [31]$$

De este modo, el parámetro β_1 asociado a *X* representa el cambio en el logaritmo de las chances al pasar la covariable de 0 a 1. Por lo tanto, el cociente de chances se obtiene simplemente exponenciando el valor de dicho parámetro:

$$\log(OR) = \beta_1 \Leftrightarrow OR = \exp(\beta_1),$$

de forma tal que el efecto aditivo en la escala *logit* (β_1) es un efecto multiplicativo de magnitud *exp*(β_1) en la escala de las chances.

Resumiendo, la diferencia entre dos *logits* equivale a la diferencia de los logaritmos de las chances, lo que a su vez es igual al logaritmo del cociente de chances. Por último, vale destacar una propiedad muy útil de los cocientes de chances, la de simetría, la que se expresa de la siguiente forma:

$$\log\left(\frac{\Pr(Y=1/X=1)\Pr(Y=0/X=0)}{\Pr(Y=0/X=1)\Pr(Y=1/X=0)}\right) = \log\left(\frac{\Pr(Y=0/X=0)\Pr(Y=1/X=1)}{\Pr(Y=1/X=0)\Pr(Y=0/X=1)}\right). \quad [32]$$

3.2.3. Enfoque marginal

En un modelo marginal, la variable respuesta es tratada separadamente de la correlación intra-*cluster*. Lo que se describe en la regresión es la esperanza marginal de *Y* como una función de las covariables, entendiendo por ella a la respuesta promedio para la subpoblación que comparte un mismo valor de las *X*. Siguiendo a Diggle *et al.* (2002), un modelo marginal supone que:

- La esperanza marginal depende de las covariables a través de la función de enlace.
- La varianza marginal depende de la media marginal de acuerdo a la función de varianza.
- La correlación entre las observaciones es función de la media marginal y de otros parámetros adicionales.

Los modelos lineales generalizados marginales permiten inferir acerca de valores promedio para toda la población, dado que $exp(\beta)$ representa un cociente de chances promedio *(population average inference)*. Si todos los individuos con el mismo nivel en las covariables tienen igual probabilidad de éxito, la frecuencia poblacional equivale a la frecuencia individual. Sin embargo, cuando hay heterogeneidad entre sujetos para el mismo nivel de las *X*, la frecuencia poblacional es el promedio de los valores individuales.

Con covariables dicotómicas, el exponencial del coeficiente estimado para cada *X* equivale a un cociente entre las chances de los subgrupos con $X=1$ y con $X=0$, manteniendo las demás covariables constantes. Si las variables regresoras son continuas, $exp(\beta)$ representa el cambio multiplicativo en las chances asociado a un aumento unitario de la covariable.

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 verosimilitud completa o condicional–.

Un supuesto implícito del modelo marginal es que la respuesta de cada miembro de un *cluster* no depende de los valores de las covariables para las otras respuestas dentro del mismo *cluster*. Por lo tanto, se admite que las covariables permanezcan constantes o que asuman distintos valores dentro de un cierto conglomerado (Pendergast *et al.*, 1996).

Ecuaciones de estimación generalizadas (GEE)

El método de estimación de cuasi-verosimilitud, originalmente desarrollado para los miembros de la familia exponencial de distribuciones, ha posibilitado difundir los modelos lineales generalizados a una gama más amplia de aplicaciones (Littell *et al.*, 1996). En lugar de asumir una distribución de probabilidad para la variable respuesta, este método sólo requiere especificar sus dos primeros momentos: *(a)* la relación entre la media y el predictor lineal; *(b)* la función de varianza.

Con respuestas binarias, e.g., puede asumirse que el *logit* de la probabilidad depende linealmente de las covariables, sin que sea necesario asumir que la distribución pertenece a la familia exponencial (Zeger *et al.*, 1988; Fahrmeir & Tutz, 2001). En tal caso, sólo deben especificarse la probabilidad de éxito y las correlaciones correspondientes al vector de respuestas binarias para cada *cluster* (Lipsitz *et al.*, 1994).

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, permitiendo que los *clusters* difieran de tamaño (Longford, 1994; Agresti, 2002).

El método de *GEE* 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* que dependen de α y β (SAS Institute Inc., 1999a) [7]:

$$S_{\beta}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right)=\sum_{i=1}^{k}\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\,\hat{V}_i^{-1}\left(y_i-\hat{\mu}_i\right)=0\,, \qquad \textbf{[33]}$$

donde:

- $\boldsymbol{\alpha}$ es un vector de parámetros asociados con un modelo específico para la correlación entre las respuestas de un mismo *cluster*.

- $\frac{\partial \hat{\mu}_i}{\partial \hat{\beta}}$ es una matriz de derivadas parciales de la media con respecto a los estimadores.
- $\hat{V}_i$ es una matriz estimada de covarianzas de trabajo, función de cierta estructura de covarianza que depende de α y β.

Si *g* representa la función de enlace, la matriz *p x* m_i –con *p* el número de parámetros y m_i el tamaño del *cluster*– de derivadas parciales de la media con respecto a los parámetros estimados, para el i-ésimo *cluster,* está dada por:

$$\left(\frac{\partial \hat{\mu}_i}{\partial \hat{\beta}}\right)' = \begin{bmatrix} \frac{x_{i11}}{g'\left(\hat{\mu}_{i1}\right)} & \cdots & \frac{x_{im_i1}}{g'\left(\hat{\mu}_{im_i}\right)} \\ \vdots & & \vdots \\ \frac{x_{i1p}}{g'\left(\hat{\mu}_{i1}\right)} & \cdots & \frac{x_{im_ip}}{g'\left(\hat{\mu}_{im_i}\right)} \end{bmatrix}. \qquad [34]$$

Liang & Zeger (1986) parametrizan los coeficientes de correlación como una función lineal de $\boldsymbol{\alpha}$:

$$\rho\left(\boldsymbol{\alpha}\right) = \alpha_{rs} \quad r \neq s. \qquad [35]$$

No obstante, la asociación entre las respuestas de un mismo *cluster* puede especificarse de distintas formas, e.g., parametrizando a V_i como una función de los cocientes de chances de a pares (Lipsitz *et al.*, 1994).

El vector $\boldsymbol{\alpha}$ suele ser desconocido y debe estimarse. Una alternativa consiste en adicionar un segundo conjunto de ecuaciones de estimación $S_\alpha(\boldsymbol{\alpha},\beta)=0$ para resolverlo simultáneamente con [33]. Otra opción es reemplazar a $\boldsymbol{\alpha}$ por un estimador consistente $\alpha(\beta)$, solución que, según Liang & Zeger (1986), es asintóticamente tan eficiente como conocer su verdadero valor. En tal caso, los parámetros de esta matriz se estiman como una función de los residuos de Pearson.

[7] Las ecuaciones son de *score* si la variable respuesta posee una distribución perteneciente a la familia exponencial.

El método de *GEE* resuelve iterativamente las ecuaciones antes expresadas, reemplazando a los parámetros de V_i por sus estimaciones. El modelo especificado por *GEE* se ajusta mediante una modificación al algoritmo del *score* de Fisher, el cual consta de los siguientes pasos (Liang & Zeger, 1986; SAS Institute Inc., 1999):

1. Hallar los estimadores iniciales de β con un modelo lineal generalizado ordinario que asuma independencia.
2. Computar la matriz de correlación de trabajo $R_i(\alpha)$, en base a los residuos de Pearson estandarizados, los valores corrientes de los parámetros y la estructura asumida para $R_i(\alpha)$. Dicha matriz, de tamaño $m_i \times m_i$, es caracterizada por el vector de parámetros α y refleja la estructura de correlación asumida.
3. Estimar la matriz de covarianzas:

$$\hat{V}_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}, \qquad [36]$$

siendo:

- A_i una matriz diagonal que contiene las funciones de varianza correspondientes a cada observación. Sus elementos están completamente especificados por las distribuciones marginales.
- ϕ el parámetro de sobredispersión.

4. Actualizar el valor de los parámetros:

$$\hat{\beta}_{r+1} = \hat{\beta}_r + \left[\sum_{i=1}^{k}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\right)'\hat{V}_i^{-1}\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\right]^{-1}\left[\sum_{i=1}^{k}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\right)'\hat{V}_i^{-1}\left(y_i-\hat{\mu}_i\right)\right]. \qquad [37]$$

5. Iterar los pasos 2 a 4 hasta alcanzar la convergencia.

Si $R_i(\alpha)$ es la verdadera matriz de correlación de Y, V_i es entonces la verdadera matriz de covarianzas de Y. La matriz de correlación de trabajo es usualmente desconocida y se la estima en el proceso iterativo de ajuste. Aún cuando $R_i(\alpha)$ puede diferir de un *cluster* a otro, los parámetros α son los mismos para la totalidad de las observaciones.

Los estimadores *GEE* de los parámetros del modelo son consistentes aún si se especifica incorrectamente la estructura de covarianza. Ello se debe a que la consistencia –

i.e., estimadores que convergen en probabilidad a los verdaderos parámetros– es una propiedad que depende del primer momento de la distribución pero no del segundo[8]. Se supone que el modelo es correcto en el sentido de que la función de enlace y el predictor lineal elegidos son adecuados, aunque la estructura de dependencia propuesta sea incorrecta (Spiess & Hamerle, 2000; McCulloch & Searle, 2001; Agresti, 2002).

La varianza de los estimadores se estima en forma robusta con la siguiente expresión, la cual recibe el nombre de matriz "s*andwich*":

$$\hat{V}_{\beta} = \left[\sum_{i=1}^{k}\left(\frac{\partial\hat{\mu}_i}{\partial\beta}\right)'\hat{V}_i^{-1}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\right)\right]^{-1}\left[\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\right)'\hat{V}_i^{-1}\left(y_i-\hat{\mu}_i\right)\left(y_i-\hat{\mu}_i\right)'\hat{V}_i^{-1}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\right)\right]\left[\sum_{i=1}^{k}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\right)'\hat{V}_i^{-1}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}}\right)\right]^{-1} \quad \textbf{[38]}$$

Dicha matriz 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. Cuanto más difiera la verdadera matriz de la matriz de correlación de trabajo que ha sido elegida, mayor es el impacto que tiene la evidencia provista por los datos (Lipsitz *et al.*, 1994; Agresti, 2002).

Dado que las estimaciones son robustas a la elección de $R_i(\alpha)$, los intervalos de confianza resultan asintóticamente correctos. No obstante, elegir una estructura de correlaciones más cercana a la verdadera redunda en un incremento de la eficiencia (Zeger & Liang, 1986). En resumen, siempre que la media esté correctamente especificada y se cumplan las condiciones usuales de regularidad, el método de *GEE* ofrece estimadores consistentes y asintóticamente normales (Zeger *et al.*, 1988):

$$\hat{\beta} \sim_a N\left(\beta, \hat{V}_{\beta}\right). \quad \textbf{[39]}$$

Con respuestas binarias, los parámetros de asociación pueden formularse y estimarse de múltiples maneras. Si se utilizan los cocientes de chances para describir la asociación, α puede estimarse mediante *ALR (Alternating Logistic Regression)*, algoritmo que alterna entre

[8] En cambio, el método alternativo *GEE2*, que adiciona ecuaciones de estimación para la estructura de correlación y conlleva una ganancia en eficiencia, tiene la desventaja de producir estimadores inconsistentes si esta parte del modelo es mal especificada (Agresti, 2002).

actualizar el modelo para la media y actualizar los logaritmos de los cocientes de chances. Alcanzada la convergencia, se obtiene una estimación de los parámetros para la media *(β)*, los parámetros para los logaritmos de los cocientes de chances *(α)*, sus errores estándares y sus covarianzas (SAS Institute Inc., 1999a).

En cuanto al parámetro de sobredispersión ϕ, éste puede estimarse luego de la última iteración. El cociente entre el estadístico X^2 de Pearson y sus correspondientes grados de libertad, ofrece un estimador consistente (Liang & Zeger, 1986):

$$\hat{\phi} = \frac{1}{n-p}\sum_{i}^{k}\sum_{j}^{m_i}\hat{r}_{ij}^{2} = \frac{1}{n-p}X^2 , \qquad [40]$$

donde:

- $\hat{r}_{ij}$ son los residuos de Pearson.
- *(n – p)* son los grados de libertad.

Modelación de la estructura de dependencia

Además de considerar los efectos de las covariables sobre la esperanza marginal, los modelos marginales requieren que se especifique una estructura de correlación para las observaciones pertenecientes a un mismo *cluster*. En forma general:

$$Var\left(y_{ij}\right) = \mu_{ij}\left(1-\mu_{ij}\right),$$

$$\text{corr}\left(y_{ir}, y_{is}\right) = \rho\left(\mu_{ir}, \mu_{is}, \boldsymbol{\alpha}\right) \qquad \forall r \neq s , \qquad [41]$$

con $\rho(\cdot)$ una función conocida. Entre las estructuras de correlación más comunes se encuentran:

- ⊡ **Independencia**: $\text{corr}\left(y_{ir}, y_{is}\right) = 0 .$ [42]
- ⊡ **Fija o conocida**: $\text{corr}\left(y_{ir}, y_{is}\right) = \alpha_{rs0} .$ [43]
- ⊡ **Intercambiable o de simetría compuesta**: $\text{corr}\left(y_{ir}, y_{is}\right) = \alpha .$ [44]

- **Sin estructura**: $\text{corr}\left(y_{ir}, y_{is}\right) = \alpha_{rs}$. **[45]**

Usualmente, no se dispone de información a priori y la correlación es considerada como ruido *(nuisance)*. La estructura más simple es la que asume a las observaciones no correlacionadas entre sí –[*42*]–[9]. En este caso, las ecuaciones de estimación generalizadas tienen la forma usual de las ecuaciones de *score* y ningún parámetro α de asociación debe estimarse conjuntamente con los β (Fahrmeir & Tutz, 2001)[10].

De acuerdo con Agresti (2002), la desventaja de elegir esta opción no se relaciona con los parámetros estimados sino con sus errores estándares. No obstante, pueden obtenerse estimaciones más robustas de los errores estándares con el método de *GEE* que incorpora la dependencia empírica que exhiben los datos. De este modo, las estimaciones basadas en el supuesto de independencia son actualizadas usando la información que otorga la muestra acerca de la verdadera estructura de dependencia.

En una primera instancia, Littell *et al.* (1996) aconsejan utilizar el modelo de correlación sin estructura –[*44*]– por ser el más general y, una vez obtenidos los resultados, elegir alguna otra estructura más parsimoniosa que se adapte al patrón exhibido por los datos. Si bien esta opción contempla una correlación distinta para cada par de observaciones de un mismo *cluster*, implica una pérdida de eficiencia ante la gran cantidad de parámetros adicionales que deben ser estimados (Agresti, 2002). Un supuesto más flexible y realista que el de independencia o falta de estructura, lo brinda la estructura de correlación de simetría compuesta –[*45*]–, la cual supone que la correlación es la misma entre cualquier par de observaciones.

Según Fahrmeir & Tutz (2001), la elección de la matriz de correlación de trabajo debe combinar la simplicidad con la pérdida de eficiencia debida a una especificación incorrecta. Si la asociación es una cuestión secundaria, ellos proponen que se privilegie un modelo sencillo, tal como el de independencia. Al respecto, Liang & Zeger (1986) mencionan que los estimadores basados en esta estructura poseen sorprendentemente buena eficiencia si la verdadera correlación es entre débil y moderada.

[9] Sólo supone independencia para datos con distribución normal.

Si las correlaciones son modestas, cualquier estructura de correlación produce estimadores *GEE* y errores estándares similares. La dependencia empírica tiene un mayor impacto al ajustar los errores estándares que ignoran la existencia de correlación. A menos que se esperen grandes diferencias, Agresti (2002) recomienda utilizar la estructura de simetría compuesta que reconoce la dependencia con el costo de un único parámetro adicional.

Un problema que surge al modelar observaciones binarias, es que los coeficientes de correlación dependen en forma compleja de las medias. La correlación entre dos respuestas binarias y_{ir} e y_{is} con medias μ_{ir} y μ_{is}, está dada por:

$$corr\left(y_{ir}, y_{is}\right) = \frac{\Pr\left(y_{ir}=1, y_{is}=1\right) - \mu_{ir}\mu_{is}}{\left[\mu_{is}\left(1-\mu_{is}\right)\mu_{ir}\left(1-\mu_{ir}\right)\right]^{1/2}}, \qquad [46]$$

donde la probabilidad conjunta $\Pr\left(y_{ir}=1, y_{is}=1\right)$ está restringida, lo cual puede limitar considerablemente el rango de correlaciones admitidas:

$$máx\left(0, \mu_{ir}+\mu_{is}-1\right) \le \Pr\left(y_{ir}=1, y_{is}=1\right) \le mín\left(\mu_{ir}, \mu_{is}\right). \qquad [47]$$

Otra forma de contemplar la dependencia con respuestas binarias es empleando el algoritmo *ALR*, que permite describir la asociación mediante los logaritmos de los cocientes de chances marginales:

$$OR\left(y_{ir}, y_{is}\right) = \frac{\Pr\left(y_{ir}=1,\ y_{is}=1\right)\ \Pr\left(y_{ir}=0,\ y_{is}=0\right)}{\Pr\left(y_{ir}=1,\ y_{is}=0\right)\ \Pr\left(y_{ir}=0,\ y_{is}=1\right)}. \qquad [48]$$

La ventaja de parametrizar la asociación en términos de cocientes de chances marginales es que ellos poseen menos restricciones que los coeficientes de correlación. Además, su interpretación no depende del tamaño de la muestra, como sucede con los cocientes de chances condicionales (Diggle *et al.*, 2002). El algoritmo *ALR* busca modelar:

$$\alpha_{irs} = \log\left(OR\left(y_{ir}, y_{is}\right)\right) = z'_{irs}\gamma, \qquad [49]$$

[10] Esto mismo sucede si se opta por la estructura conocida –[*43*]–, en donde se fijan valores para los coeficientes de

donde:

- γ es un vector de parámetros de regresión.
- z'_{irs} es un vector fijo de coeficientes.

Los logaritmos de los cocientes de chances (α_{irs}) modelados de esta manera pueden asumir distintos valores en los subgrupos definidos por z'_{irs} –e.g., subgrupos dentro de *clusters* o efectos de bloque entre *clusters*–. Especificar un modelo de regresión para los logaritmos de los cocientes de chances requiere definir las filas de la matriz *Z* para cada *cluster i* y para cada par *(r, s)*. Algunos de los métodos disponibles para especificar dicha matriz, junto con el comando para invocarlos en SAS PROC GENMOD, son (SAS Institute Inc., 1999a):

- **Logaritmos de cocientes de chances intercambiables** (LOGOR=EXCH): los logaritmos de los cocientes de chances permanecen constantes para todos los *clusters i* y todos los pares *(r, s)*, siendo el parámetro α el logaritmo del cociente de chances común. Este es el modelo más simple, análogo al supuesto de equicorrelación. Al parametrizar a los logaritmos de los cocientes de chances de esta forma se reduce el número de estimaciones a realizar (Fahrmeir & Tutz, 2001).
- **Logaritmos de cocientes de chances por *cluster*** (LOGOR=LOGORVAR): debe especificarse una variable argumento que define el efecto de bloque entre *clusters*. Los logaritmos de los cocientes de chances son constantes al interior del *cluster* pero asumen distintos valores en cada nivel de la variable especificada.
- **Logaritmos de cocientes de chances anidados a un nivel** (LOGOR=NEST1): debe especificarse la variable que define los *subclusters*. Existen dos parámetros para los logaritmos de los cocientes de chances en este modelo: uno correspondiente a los pares del mismo *cluster* pero distinto *subcluster* y otro correspondiente a los pares del mismo *cluster* y *subcluster*.
- **Logaritmos de cocientes de chances anidados a *k* niveles** (LOGOR=NESTK): deben especificarse las distintas variables que definen los *subclusters*. Dentro de

correlación intra-*cluster*.

cada *cluster* se computa un parámetro para los pares que poseen el mismo valor de la variable para ambos miembros y otro parámetro para cada combinación única de diferentes valores de la variable.

- ***Clusters* completamente parametrizados** (LOGOR=FULLCLUST): cada *cluster* es parametrizado de la misma manera y existe un parámetro para cada par único dentro del *cluster*.

Siempre que el interés resida en estimar los coeficientes de regresión, el esfuerzo debe dirigirse a modelar la media marginal usando una aproximación razonable de la correlación. La robustez de la inferencia acerca de los parámetros puede juzgarse ajustando el mismo modelo con distintas estructuras de dependencia y comparando los estimadores y sus errores estándares. Sólo si ellos difieren sustancialmente será necesario un tratamiento más cuidadoso del tema (Diggle *et al.*, 2002).

3.2.4. Enfoque mixto

Siguiendo a Diggle *et al.* (2002), un modelo mixto constituye una descripción razonable si el conjunto de coeficientes de una población de individuos puede considerarse como una muestra aleatoria a partir de una cierta distribución. Dados los efectos aleatorios *(U_i)*, puede asumirse que las observaciones en el i-ésimo *cluster* son independientes entre sí y que corresponden a un modelo lineal generalizado con densidad perteneciente a la familia exponencial[11]:

$$f(Y_{ij} \;/\; U_i,\, \phi) = \exp\left[\sum_{i=1}^{k}\sum_{j=1}^{m_i} y_{ij}\;\theta_{ij} \;-\; \sum_{i=1}^{k}\sum_{j=1}^{m_i}\frac{b_{ij}(\theta_{ij})}{a(\phi)} \;+\; c_i\;(y_{ij},\;\phi)\right]. \qquad [50]$$

Bajo el enfoque mixto, puede optarse por formular un modelo con verosimilitud completa o con verosimilitud condicional. Seguidamente se presentan ambas alternativas.

[11] Siempre que, condicional a los efectos aleatorios, los componentes del vector de respuestas se distribuyan independientemente y su distribución sea un miembro de la familia exponencial (Schall, 1991).

Modelo mixto con verosimilitud completa

Este modelo parte de suponer que los coeficientes para cada *cluster* son de interés. Se predicen los efectos aleatorios *(U_i)* y se estiman los efectos fijos *(β)* al integrar o promediar sobre los efectos aleatorios, siendo la verosimilitud completa adecuada para la estimación del modelo.

La idea básica en los modelos mixtos es la existencia de una heterogeneidad natural entre los coeficientes de regresión de los *clusters*, e.g. en los interceptos, que puede representarse mediante una distribución de probabilidad. La correlación entre observaciones de un mismo conglomerado implica que los individuos que a él pertenecen comparten variables no observables U_i. La distribución de probabilidad de U_i típicamente se considera normal con media cero y varianza desconocida[12]:

$$U_i \sim N(0, G), \qquad [51]$$

lo que da lugar al modelo logístico-normal:

$$g\left[E\left(y_{ij}/U_i\right)\right] = \text{logit}\ \Pr\left(y_{ij}=1/U_i\right) = \eta_{ij} = x_{ij}'\beta + U_i. \qquad [52]$$

Bajo la formulación de intercepto aleatorio, los valores realizados de U_i son una cantidad por la cual todas las mediciones del conglomerado en cuestión se ven incrementadas o disminuidas con relación a un *cluster* típico. Los parámetros se estiman en forma consistente bajo el supuesto que establece que los efectos aleatorios son independientes de las covariables. Las covariables pueden variar de un *cluster* a otro, manteniéndose constantes para todos los miembros de un mismo conglomerado *(between-cluster)*, o asumir distintos valores para los individuos de un mismo conglomerado *(within-cluster)*.

Los parámetros estimados representan los efectos de las covariables sobre las chances de un *cluster* particular, por lo que se obtienen coeficientes específicos para cada *cluster (cluster-specific inference)*. Es decir, los coeficientes de regresión describen la respuesta de

[12] La suposición habitual acerca de la distribución normal de los efectos aleatorios se explica porque suele ser difícil justificar una distribución particular (Schall, 1991).

cada *cluster* ante cambios en el nivel de las covariables, estimándose el cambio esperado en las probabilidades individuales (Zeger *et al.*, 1988).

Aún cuando este modelo es asemejable al modelo lineal mixto, los componentes de varianza no se pueden estimar por máxima-verosimilitud restringida *(Restricted Maximum Likelihood - REML)*. El análogo de este método es difícil de implementar al incluir integrales de altas dimensiones sobre la totalidad de los efectos fijos, las cuales no suelen obtenerse en forma cerrada. Siguiendo a Diggle *et al.* (2002), si se trata a U_i como una muestra independiente a partir de una distribución normal de efectos aleatorios, la función de verosimilitud para β y G es proporcional a:

$$L(\boldsymbol{\beta},G;\mathbf{y}) \propto \prod_{i=1}^{k}\int\prod_{j=1}^{m_i}\left\{\mu_{ij}(\beta,U_i)\right\}^{y_{ij}}\left\{1-\mu_{ij}(\beta,U_i)\right\}^{1-y_{ij}} f(U_i,G)\,dU_i\,, \qquad [53]$$

$$\mu_{ij}(\beta,U_i) = E(y_{ij}\,/\,U_i;\beta). \qquad [54]$$

Con el enlace logístico, el supuesto acerca de la distribución normal de U_i y la formulación de intercepto aleatorio, la función de verosimilitud se expresa como:

$$\prod_{i=1}^{k}\int \exp\left[\beta'\sum_{j=1}^{m_i}x_{ij}y_{ij} + U_i'\sum_{j=1}^{m_i}y_{ij} - \sum_{j=1}^{m_i}\log\left\{1+\exp\left(x_{ij}'\beta+U_i\right)\right\}\right](2\pi)^{-1}|G|^{-q/2}\exp\left(\frac{-U_i'G^{-1}U_i}{2}\right)dU_i\,, \qquad [55]$$

donde G es una matriz de covarianzas $p \times p$ para cada U_i. Lamentablemente, esta expresión es intratable en forma analítica. No obstante, la integración numérica puede lograrse con distintos enfoques. SAS PROC NLMIXED utiliza la cuadratura adaptativa gausiana *(Adaptive Gaussian Quadrature)* como método de integración[13]. Éste intenta hallar los nodos y ponderaciones que minimizan el error cometido al sustituir el valor real de la integral, por el valor aproximado mediante una suma ponderada sobre abscisas predefinidas para los efectos aleatorios.

El método de optimización empleado por SAS PROC NLMIXED es el algoritmo dual cuasi-Newton, el cual utiliza el gradiente y no requiere el cómputo de derivadas de segundo orden. Este algoritmo calcula en la k-ésima iteración un escalar α^k que optimiza una función

no lineal *f(α)* en una dirección descendente de búsqueda, la que a su vez trata de optimizar una aproximación lineal o cuadrática de la función objetivo *f(x)* de *p* parámetros (SAS Institute Inc., 1999a).

Modelo mixto con verosimilitud condicional

Cuando el interés recae en un subconjunto de coeficientes de regresión, ninguno de los cuales varía entre *clusters*, las variables U_i pueden considerarse como auxiliares *(nuisance)* –i.e., como si fueran parámetros fijos– y condicionarse fuera del modelo. La verosimilitud condicional puede utilizarse entonces como método de estimación[14].

Una razón que respalda esta estrategia de modelación es que, si existe dependencia, las distribuciones condicionales son conceptualmente más sencillas. En este caso, los parámetros se interpretan en términos de probabilidades condicionales (Pendergast *et al.*, 1996). Bajo este enfoque, el modelo más simple plantea que cada individuo está sujeto a la influencia de variables que, si bien no pueden medirse, afectan por igual a todos los miembros de un mismo *cluster*:

$$\eta_{ij} = \text{logit}\,\Pr\left(y_{ij} = 1 \,/\, U_i\right) = \beta_0 + \beta\, x_{ij} + U_i = \left(\beta_0 + U_i\right) + \beta\, x_{ij} = \gamma_i + \beta\, x_{ij}\,. \qquad [56]$$

Con el modelo condicional y la formulación de intercepto aleatorio, la inferencia resulta consistente aún si no se cumple el supuesto de que los efectos aleatorios sean independientes de las covariables. Siguiendo a Diggle *et al.* (2002), a partir del modelo de regresión logística y asumiendo que el vector de las *X* no incluye al intercepto, la función de verosimilitud conjunta para β y γ_i es proporcional a:

$$\prod_{i=1}^{k} \exp\left[\gamma_i \sum_{j=1}^{m_i} y_{ij} + \left(\sum_{j=1}^{m_i} y_{ij}\, x'_{ij}\right)\beta - \sum_{j=1}^{m_i} \log\left\{1 + \exp\left(\gamma_i + x'_{ij}\beta\right)\right\}\right]. \qquad [57]$$

[13] Los métodos de cuadratura son técnicas de integración numérica que aproximan integrales que no pueden resolverse analíticamente.
[14] Operar en términos condicionales es una estrategia adecuada siempre que se abandona el supuesto de independencia. Se dice que un modelo está especificado condicionalmente si la distribución conjunta de los datos se construye a partir de las distribuciones condicionales (Pendergast *et al.*, 1996).

La función de verosimilitud condicional para β, dado el estadístico suficiente para γ_i, queda expresada como:

$$L\left[\beta / \sum y_{ij}\right] = \prod_{i=1}^{k} \frac{\exp\left(\sum_{j=1}^{m_i} y_{ij} x'_{ij}\beta\right)}{\sum_{R_i} \exp\left(\sum_{\ell=1}^{y_{i\cdot}} x'_{i\ell}\beta\right)}, \qquad [58]$$

donde R_i contiene todas las formas posibles de obtener $y_{i\cdot}$ "éxitos" en m_i observaciones dentro de un *cluster*. Esta expresión resulta similar a la función de verosimilitud parcial del modelo de riesgo proporcional de Cox utilizada en el análisis de sobrevida, lo que permite que cualquier *software* que pueda ser usado en estos casos también ajuste el modelo condicional con intercepto aleatorio (SAS PROC PHREG).

La principal ventaja del enfoque condicional es que se eliminan los efectos aleatorios de la función de verosimilitud, sin necesidad de asumir que ellos provienen de una determinada distribución de probabilidad. Como desventaja, debe mencionarse que se confía totalmente en las comparaciones intra-*cluster*: cualquier *cluster* con $y_{i\cdot}=m_i$ o $y_{i\cdot}=0$ no da información alguna sobre los coeficientes de regresión. Como consecuencia, los errores estándares de los estimadores tienden a ser mayores que bajo un análisis marginal o mixto con verosimilitud completa.

3.2.5. Relación entre los enfoques marginal y mixto

Si la variable respuesta es binaria, optar entre un modelo lineal generalizado marginal o mixto tiene consecuencias sobre la interpretación de los coeficientes. Como las respuestas procedentes de distintos *clusters* se asumen independientes entre sí, los modelos marginales describen el efecto de las covariables sobre las respuestas promedio, lo cual contrasta con el modelo mixto en el que los efectos varían de un *cluster* a otro (Fahrmeir & Tutz, 2001).

Algunas de las diferencias entre ambos enfoques se esquematizan en la Tabla 1:

TABLA 1: Diferencias entre los enfoques marginal y mixto

Enfoque marginal	Enfoque mixto
✓ Se describe la $Pr(y_{ij} = 1)$ como función de las covariables.	✓ Se describe la $Pr(y_{ij} = 1 / U_i)$ como función de las covariables.
✓ $exp(\beta)$ representa un cociente de chances poblacional.	✓ $exp(\beta)$ representa un cociente de chances para *clusters* particulares.
✓ β mide el cambio en el *logit* de la probabilidad ante un incremento unitario en el valor de *X*, controlando por las restantes covariables.	✓ β mide el cambio en el *logit* condicional de la probabilidad ante un incremento unitario en el valor de *X* para los individuos pertenecientes al mismo *cluster*, controlando por las restantes covariables.

Ambas estrategias pueden distinguirse si se tiene en cuenta que con el modelo marginal se describe la distribución marginal y los parámetros estimados representan un vector de respuestas típico. En tanto, el modelo mixto se concentra en modelar a los *clusters* con el fin de comprender a la población; sus estimadores se refieren a comportamientos individuales, obteniéndose parámetros típicos para los efectos fijos (Lindstrom & Bates, 1990).

Una ventaja del enfoque marginal es que la respuesta promedio poblacional para un valor dado de *X* es directamente estimable a partir de las observaciones, sin necesidad de efectuar supuesto alguno acerca de la heterogeneidad de los parámetros a través de los *clusters*. Si bien dichos parámetros se hallan más próximos a los datos que los correspondientes a un modelo mixto, aún dependen del grado de heterogeneidad existente en la población (Zeger *et al.*, 1988).

Otra diferencia a resaltar es que para que la inferencia acerca de los parámetros sea consistente, el enfoque marginal sólo requiere que la función de enlace sea correctamente especificada. En cambio, el modelo mixto con verosimilitud completa exige que la función de enlace y la distribución de probabilidad asumida para los efectos aleatorios se especifiquen correctamente, así como que dichos efectos sean independientes de las covariables.

Bajo el enfoque marginal se modelan por separado los efectos de las covariables sobre la variable respuesta y la asociación entre las respuestas. Ello contrasta con el enfoque condicional, el cual está completamente definido por la estructura condicional de medias. Por tal motivo, éste último resulta de utilidad si las distribuciones condicionales revisten interés –e.g., con fines predictivos–, pero si el objetivo es analizar los efectos de las covariables sobre la variable respuesta, los modelos marginales son más adecuados (Fahrmeir & Tutz, 2001).

Asimismo, la interpretación provista por el enfoque condicional se dificulta si las covariables se mantienen constantes dentro de los *clusters*, debido a que el método aprovecha la información provista cuando los *clusters* actúan como control y se observa el cambio de las covariables al interior de los mismos. En tal caso, el modelo mixto con verosimilitud completa provee una interpretación más satisfactoria (Neuhaus *et al.*, 1991).

La dificultad computacional, inherente a la estimación de los modelos mixtos, ha contribuido a la popularidad del modelo marginal para datos binarios provenientes de conglomerados. Al respecto, Neuhaus *et al.* (1991) destacan que las pruebas tipo Wald tienden a poseer niveles de significatividad muy similares en ambos modelos, ya que los coeficientes asociados a los efectos fijos son prácticamente iguales cuando se los estandariza. No obstante, si el modelo mixto es correcto, contrastar la hipótesis nula $\beta=0$ usando el enfoque marginal, tendrá el nivel correcto de error de tipo I pero será menos eficiente.

3.2.6. Inferencia y diagnóstico del modelo

Contrastes de hipótesis e intervalos de confianza

En el contexto de los modelos lineales generalizados se utilizan tres métodos para contrastar hipótesis lineales acerca de los parámetros: Wald, cociente de verosimilitud y *score*. Aunque asintóticamente son equivalentes, los resultados suelen diferir cuando las muestras no son demasiado grandes. Una marcada divergencia entre los valores también es indicio de falta de normalidad en la distribución de los estimadores.

Las pruebas basadas en la función de verosimilitud son adecuadas cuando se emplean modelos mixtos, pero no si se opta por el enfoque marginal (Fahrmeir & Tutz, 2001). Puesto que el método de *GEE* no especifica completamente la distribución conjunta, carece de una función de verosimilitud. Por consiguiente, no son válidos los métodos basados en el cociente de verosimilitud para evaluar la bondad del ajuste, comparar modelos o efectuar inferencia acerca de los parámetros.

En este caso, la inferencia puede basarse en los estadísticos de Wald o de cuasi-*score.* Estos últimos –iguales a la derivada de la función de cuasi-verosimilitud– resultan superiores, siendo sus propiedades asintóticas similares a las de las pruebas de *score* y, por lo tanto, más conservadores que las pruebas de Wald para tamaños de muestra finitos.

- **Estadístico de *Wald***:

$$z_{Wald} = \left(\frac{\hat{\beta}}{ASE\left(\hat{\beta}\right)} \right) \sim N(0,1) \qquad [59]$$

El estadístico de Wald tiene una distribución aproximadamente normal estándar bajo la hipótesis nula (H_0) que postula que $\beta=0$. De forma equivalente, para hipótesis alternativas (H_a) bilaterales, z^2 posee, bajo la hipótesis nula, una distribución aproximada chi-cuadrado con un grado de libertad. Este tipo de estadístico utiliza el error estándar asintótico *(Asymptotic Standard Error - ASE)* no nulo –i.e., evaluado en $\hat{\beta}$–, el cual se calcula a partir de la curvatura de la función de log-verosimilitud en su máximo.

La forma multivariada de la prueba, tiene como estadístico:

$$W = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)' \left[\text{cov}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right), \qquad [60]$$

donde la covarianza no nula se basa en la curvatura de la función de log-verosimilitud evaluada en el valor estimado. La distribución normal multivariada asintótica para los estimadores implica una distribución asintótica chi-cuadrado para *W*, con grados de libertad iguales al rango de la matriz de covarianzas (Agresti, 2002).

- **Estadístico del *cociente de verosimilitud (likelihood-ratio)***:

$$\chi^2_{LR} = -2\log\left(\frac{\ell_0}{\ell_1}\right) = -2\left[\log(\ell_0) - \log(\ell_1)\right] = -2(L_0 - L_1) \sim \chi^2_q \qquad [61]$$

L_0 y L_1 representan las funciones de log-verosimilitud maximizadas sobre un conjunto posible de valores bajo la hipótesis nula y sobre un conjunto de parámetros más amplio (H_a), respectivamente. Una gran diferencia entre ambas conduce a rechazar la hipótesis formulada. Para muestras grandes, el estadístico tiene una distribución chi-cuadrado con grados de

libertad equivalentes a la diferencia en dimensiones de los espacios paramétricos bajo H_0 y H_a (Agresti, 2002).

- **Estadístico de *score***:

$$\chi^2_{\text{Score}} = \left(\frac{L_0'}{ASE\left(L_0'\right)} \right)^2 \sim \chi^2_q \qquad [62]$$

La función de *score* equivale a la primera derivada de la función de log-verosimilitud con respecto a los parámetros de interés. Este estadístico se basa en la pendiente y curvatura esperada de la función de log-verosimilitud evaluada en la hipótesis nula. Si la derivada en el valor postulado es lo suficientemente cercana a cero, ello indica, partiendo de la concavidad de la función de verosimilitud, que la log-verosimilitud bajo H_0 se encuentra cerca del máximo. En el caso multiparamétrico, el estadístico de *score* es una forma cuadrática basada en el vector de derivadas parciales de la función de log-verosimilitud con respecto a los parámetros y la inversa de la matriz de información, ambas evaluadas en la hipótesis nula (Agresti, 2002).

Esta prueba es más simple que la del cociente de verosimilitud, ya que sólo requiere que se obtengan estimaciones de los parámetros bajo la hipótesis nula. Por tal motivo, su uso ha sido propuesto en el contexto de los modelos lineales generalizados con diversos fines, tales como probar la existencia de sobredispersión o la inclusión de términos adicionales en el modelo (Smyth, 2003). La aplicación de la prueba de *score* resulta conveniente si desea compararse un modelo ajustado con otro más complejo, como es el caso del procedimiento de selección hacia delante o *forward.*

La Figura 1 es un gráfico correspondiente al caso univariado en el que se observa la información utilizada por las tres pruebas comentadas. La prueba de Wald trata de comparar si la diferencia en el eje horizontal entre el valor estimado y el asumido bajo la hipótesis nula es lo suficientemente pequeña. El cociente de verosimilitud equivale a dos veces la distancia vertical entre los valores de $L(\beta)$ evaluada en $\hat{\beta}$ y en 0; en este sentido, es la prueba que usa la mayor cantidad de información y, por lo tanto, la más versátil de las tres. Por último, la prueba de *score* juzga si la pendiente en L_0 es lo suficientemente próxima a cero.

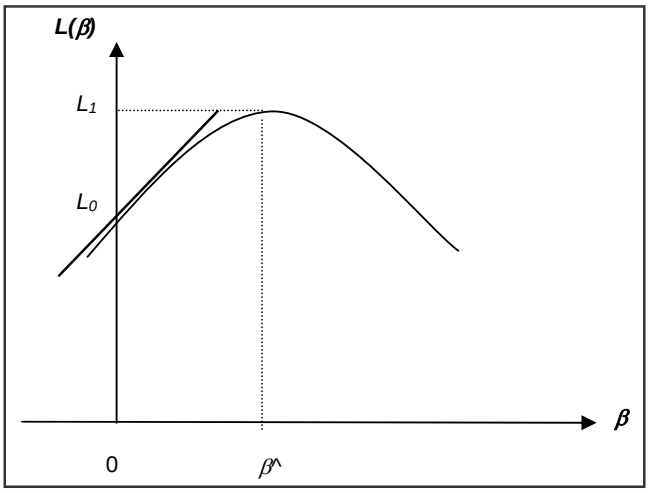

FIGURA 1: Información utilizada por las pruebas de Wald, del cociente de verosimilitud y de *score* (Fuente: Agresti, 1996)

La prueba de Wald es la más sencilla computacionalmente, pero la menos recomendable por ser excesivamente liberal cuando el tamaño de la muestra es pequeño y por no ser invariante ante cambios en la parametrización. Cuando se desea evaluar una hipótesis acerca de un parámetro binomial, el estadístico de *score* se prefiere a la prueba de Wald por utilizar el error estándar evaluado en H_0 en vez del estimado; además, su distribución nula en el muestreo se halla más próxima a la normal estándar.

En la práctica, es más informativo construir intervalos de confianza para los parámetros que contrastar hipótesis acerca de sus valores. Para las tres pruebas mencionadas, los intervalos de confianza se obtienen directamente invirtiendo la prueba, determinando así el rango de valores plausibles para los parámetros de interés (Agresti, 2002).

Los contrastes permiten evaluar hipótesis lineales específicas acerca de los parámetros del modelo, definidas en una matriz $L\beta=0$[15]. Los contrastes pueden realizarse mediante los tres métodos antes referidos, i.e., Wald, cociente de verosimilitud o cuasi-*score*, con grados de libertad determinados por el número de filas linealmente independientes de la matriz *L*. Si el análisis efectuado es de tipo III, los resultados obtenidos no dependen del orden en el cual se especifiquen los términos en el modelo (SAS Institute Inc., 1999a).

[15] Una función $L\beta$ es estimable si existe una combinación lineal de *Y* tal que su valor esperado sea $L\beta$, o si existe una combinación de las filas de *X* igual a *L* (SAS Institute Inc., 1999b).

Medidas usuales de bondad del ajuste

Luego de ajustar un modelo a un conjunto de datos, es preciso saber hasta qué punto los valores ajustados se asemejan a los valores observados, lo que conduce a efectuar un análisis de bondad del ajuste *(goodness of fit)*. Si bien se hace hincapié en aquellas técnicas que resultan adecuadas cuando las observaciones binarias se encuentran correlacionadas, cabe mencionar brevemente las medidas usuales.

La *deviance* es un estadístico de bondad del ajuste que compara la máxima verosimilitud alcanzable al ajustar el modelo bajo análisis con la obtenible con el modelo saturado, el cual no provee ninguna reducción pero resulta útil como base para la comparación[16]. Sin embargo, ésta no es una buena medida cuando se modelan observaciones binarias. En tal caso, la *deviance* depende de las observaciones a través de las probabilidades ajustadas y, por consiguiente, no aporta información alguna acerca de la similitud entre los valores ajustados y observados. En esta situación, la *deviance* es, en numerosas ocasiones, inferior a *(n–p)* que es su valor esperado.

Cuando se trabaja con observaciones binarias desagregadas –y no con la proporción de éxitos–, el estadístico de *deviance* no tiene ni siquiera una distribución aproximada chi-cuadrado. Ello se debe a que la teoría asintótica usual, la cual justifica la distribución aproximada chi-cuadrado de la *deviance*, requiere que tanto el modelo nulo como el alternativo permanezcan constantes al incrementarse el número de observaciones. Pero al tratarse de observaciones binarias, cuando su número aumenta también lo hace la cantidad de parámetros necesarios para el modelo alternativo. Por consiguiente, la *deviance* nunca es informativa acerca de la bondad del ajuste del modelo, siendo este resultado estrictamente válido si se opta por el enlace logístico (Swan, 1986).

El estadístico X^2 de Pearson, otra de las medidas habitualmente utilizadas, se define como:

[16] El modelo saturado incluye igual número de parámetros que de observaciones. Si bien cada parámetro es exacto, pues describe perfectamente la ubicación del dato, no provee ninguna reducción y posee un valor inferencial limitado. El modelo saturado se utiliza para calcular la *deviance* residual, la cual se emplea en el contexto de los modelos lineales generalizados para juzgar la bondad del ajuste (Gill, 2001).

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\left(y_{ij} - \hat{\mu}_{ij}\right)^2}{V\left(\hat{\mu}_{ij}\right)}, \qquad [63]$$

el cual resulta más robusto que la *deviance* ante una mala especificación de la distribución de la variable respuesta, al depender su valor esperado sólo de los dos primeros momentos de la función de distribución. En sí, puede considerarse como un estadístico tipo *score*, el cual contrasta el modelo ajustado versus el modelo saturado (Smyth, 2003).

Cuando la variable respuesta es discreta, X^2 converge a una distribución chi-cuadrado más rápidamente que la *deviance*. Pero, desafortunadamente, tiene la desventaja de no ser aditivo para modelos anidados y de exhibir un comportamiento pobre si el tamaño de muestra es relativamente pequeño.

Selección y diagnóstico del modelo

La adecuación de un modelo mixto puede evaluarse con los criterios del cociente de verosimilitud (LR) para modelos anidados –definido en [*61*]–, de Akaike (AIC) y de Schwarz (BIC o SBC), optándose por aquel modelo que presente el menor valor para cualquiera de los estadísticos analizados:

$$\text{AIC} = -2\ell\left(\hat{\beta}\right) + 2p, \qquad [64]$$

$$\text{BIC o SBC} = -2\ell\left(\hat{\beta}\right) + p\log(n), \qquad [65]$$

siendo:

- p el número de parámetros.
- n el número de observaciones.

AIC *(Akaike Information Criterion)* se apoya en el principio de seleccionar un modelo que minimice la negativa de la función de verosimilitud, penalizado por el número de parámetros. Este criterio es muy utilizado, aún cuando suele estar sesgado hacia modelos que sobreajustan y poseen parámetros extras. A pesar de que el término de penalización se

incrementa linealmente con el número de covariables, la función de log-verosimilitud lo hace más rápidamente. Un beneficio sustancial de la inclusión de dicho término consiste en reconocer que no es una buena estrategia basar la decisión acerca del ajuste del modelo sólo en el valor que asume la función de log-verosimilitud, ya que la misma nunca decrece al adicionar covariables en el predictor lineal.

En cuanto al BIC *(Bayesian Information Criterion)*, aún cuando surge de una perspectiva totalmente distinta, funcionalmente se asemeja al criterio de Akaike. Al contemplar el tamaño de la muestra *(n)* en su cálculo, resulta una medida más apropiada cuando se comparan modelos entre los cuales difiere el tamaño muestral. Mientras AIC tiende a favorecer la inclusión de más covariables y un mejor ajuste, BIC favorece el uso de menos covariables y un ajuste inferior (Gill, 2001)[17].

También la función de *deviance* resulta muy útil en la comparación de dos modelos anidados, e.g., cuando desea evaluarse si la adición de una nueva covariable mejora significativamente el ajuste. Dada una secuencia de modelos anidados, se utiliza la *deviance* como una medida de discrepancia entre modelos bajo la forma de una tabla de diferencias de *deviance.* Sin embargo, al procesar datos binarios no hay ortogonalidad y no es posible, por lo tanto, estimar los efectos independientemente. Por consiguiente, el orden en que los términos se agregan al modelo no es una cuestión irrelevante (Collett, 1991).

Si se estima un modelo marginal mediante el método de ecuaciones de estimación generalizadas –el cual emplea el enfoque de cuasi-verosimilitud–, los estadísticos que se basan en la función de verosimilitud (LR, AIC, BIC) no resultan válidos. Tampoco existen otros métodos disponibles que se hayan aplicado a casos concretos y que permitan seleccionar un subconjunto de covariables (Ver Anexo C).

En lo que al diagnóstico del modelo se refiere, pueden mencionarse diversos residuos usualmente empleados con tales fines. Los residuos ordinarios carecen de utilidad si la variable *Y* no se distribuye normalmente:

$$R_{ordinarios} = y_{ij} - \hat{\mu}_{ij} \,. \qquad \textbf{[66]}$$

[17] Una simulación realizada por Amemiya (1980) proporciona evidencia de que, cuando el tamaño de muestra es pequeño, BIC encuentra el modelo correcto más frecuentemente que AIC (Gill, 2001).

Los residuos de Pearson consisten en la raíz cuadrada de la i-ésima contribución al estadístico X^2 de Pearson, equivalentes a los residuos ordinarios ponderados por el desvío estándar de la predicción:

$$R_{Pearson} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{V\left(\hat{\mu}_{ij}\right)}}. \qquad [67]$$

Si bien esta formulación intenta proveer cierto sentido de escala a los residuos ordinarios, sólo en situaciones ideales y asintóticas poseen distribución normal. Lamentablemente, suelen ser asimétricos debido a su propiedad de relación media-varianza, por lo cual resultan poco útiles para variables binarias.

Los residuos de *deviance* se definen como la raíz cuadrada de la contribución de la i-ésima observación a la *deviance*:

$$R_{deviance} = signo\left[\left(y_{ij} - \hat{\mu}_{ij}\right)\right]\sqrt{d_{ij}}\ , \qquad [68]$$

pero, al no ser la *deviance* una medida válida cuando las observaciones son binarias, tampoco son útiles como medida de discrepancia en modelos para respuestas binarias.

Un tipo de residuo ocasionalmente útil es el de trabajo, el que procede del último paso del proceso iterativo de estimación. A pesar de que puede utilizarse como una medida de diagnóstico para evaluar la convergencia y el ajuste del modelo, la falta de una teoría general limita su uso. Estos residuos se definen como:

$$R_{trabajo} = \left(y_{ij} - \hat{\mu}_{ij}\right)\frac{\partial}{\partial \eta}\hat{\mu}. \qquad [69]$$

Los residuos de Anscombe, por su parte, compensan la falta de normalidad de los residuos de Pearson mediante una transformación que mitiga la asimetría asintótica de primer orden y da como resultado una distribución aproximadamente unimodal y simétrica. La función *A(y)*:

$$A\left(y_{ij}\right) = \int Var\left(\mu_{ij}\right)^{-1/3} d\mu\ , \qquad [70]$$

se aplica a la variable respuesta y a su media, ajustando los residuos a la escala de la varianza. Ellos son normalizados de la siguiente manera:

$$\frac{\partial}{\partial \hat{\mu}} A\left(y_{ij}\right)\sqrt{Var\left(\mu_{ij}\right)}\,. \qquad [71]$$

Si la variable respuesta es binomial, la aplicación directa de [*71*] conduce a la función beta incompleta, expresión que resulta intratable analíticamente. Una alternativa es utilizar la transformación de Cox & Snell (1968) que divide a la función beta incompleta por la función beta completa, obteniéndose así una forma simétrica más fácil de tabular[18]:

$$\phi(x) = \int_0^x t^{-1/3}\left(1-t\right)^{-1/3} dt = I_x\left(2/3,2/3\right)B_x\left(2/3,2/3\right). \qquad [72]$$

A pesar de que los residuos de Anscombe surgen de una forma completamente distinta a los residuos de *deviance*, ambos exhiben un comportamiento similar y logran producir estructuras aproximadamente normales. Éstos se expresan como:

$$R_{Anscombe} = \frac{\left(\phi\left(y_{ij}\right) - \phi\left(\hat{\mu}_{ij}\right)\right)}{\hat{\mu}_{ij}^{\,1/6}\left(1-\hat{\mu}_{ij}\right)^{1/6}}\,. \qquad [73]$$

También los gráficos de residuos desempeñan un rol central en el diagnóstico de los modelos estadísticos. Cuando la variable de interés es binomial, puede obtenerse un gráfico de residuos con la siguiente función del valor predicho en el eje de abscisas (Gilchrist & Green, 1996):

$$f\left(\hat{\mu}_{ij}\right) = 2sen^{-1}\sqrt{\frac{\hat{\mu}_{ij}}{m_i}}\,. \qquad [74]$$

Ninguna desviación sistemática debe ser observada; si los residuos exhiben un patrón que crece o decrece con los valores ajustados, debe sospecharse la existencia de alguna falla en el modelo –e.g., una incorrecta especificación de la función de varianza o la omisión

[18] La tabla incluida en el artículo de Cox & Snell (1968) ofrece los valores de la distribución beta incompleta $I_x(2/3, 2/3)$, simétrica en torno a 0.5.

de alguna covariable relevante–. Por otro lado, si sólo unos pocos residuos se alejan del resto, las respectivas observaciones pueden ser atípicas y merecen un examen más cuidadoso (Fahrmeir & Tutz, 2001). No obstante, es poco probable que un gráfico de este tipo resulte útil si la variable es de naturaleza binaria (Gilchrist & Green, 1996).

Otro gráfico utilizado es el *Q-Q plot (quantile versus quantile)*, con el cual es posible evaluar si la distribución de los residuos obtenidos es aproximadamente normal, dando una fuerte desviación lugar a un curva sigmoidea. Cuando lo que se desea es seleccionar un modelo entre varios, el más adecuado será aquel cuyos residuos posean una distribución más semejante a la normal[19]. Nuevamente, si la variable respuesta es binaria y hay muchos ceros, este gráfico se verá distorsionado al observarse numerosos residuos cerca del origen (Gilchrist & Green, 1996).

En este trabajo se opta por los residuos de Anscombe para el diagnóstico de los modelos ajustados. La naturaleza binaria de las observaciones, la ausencia de una función de verosimilitud bajo el enfoque marginal y la falta de implementación de algunos desarrollos recientes –comentados en el Anexo C–, dificultan notoriamente la tarea de aplicar otros residuos como elementos de diagnóstico.

[19] Si bien la normalidad de las perturbaciones es una exigencia del modelo lineal clásico, en este contexto más que una condición es una descripción útil del comportamiento de los residuos (Gill, 2001).

APLICACIÓN DE LOS MÉTODOS DE INFERENCIA

- Metodología de trabajo
- Descripción de la aplicación y diseño muestral
- Inferencia basada en diseño muestral
- Inferencia basada en modelos
  - Formulación
  - Estimación
  - Inferencia
  - Diagnóstico
  - Poder predictivo
  - Interpretación de coeficientes
  - Resumen de resultados
- Comparación entre métodos de inferencia

4. METODOLOGÍA DE TRABAJO

La metodología de trabajo que se ha seguido consta de los siguientes pasos:

- Diseño e implementación de una encuesta, con el propósito de inferir sobre la proporción de alumnos universitarios con vocación emprendedora.

- Análisis de las encuestas bajo los métodos de inferencia presentados:

  - ✓ Estimación mediante la inferencia clásica y realización de pruebas de hipótesis para evaluar la igualdad entre las proporciones de alumnos con vocación emprendedora para distintas subpoblaciones.

  - ✓ Formulación de modelos marginales con distintas estructuras de dependencia y modelos mixtos basados en la función de verosimilitud completa y condicional para modelar la vocación emprendedora. La variable respuesta se explica en función de distintas covariables y se considera la posible dependencia entre las mediciones de los individuos de un mismo *cluster*.

- Evaluación de la calidad de ajuste de los modelos, tanto predictiva como posdictiva.

A continuación se presenta el caso de estudio al que serán aplicados los citados métodos de inferencia y se describe el diseño muestral llevado a cabo. Posteriormente, se efectúa el análisis bajo ambas estrategias, las que luego son comparadas para, a partir de ello, concluir acerca de las ventajas y desventajas de cada una.

5. DESCRIPCIÓN DE LA APLICACIÓN Y DISEÑO MUESTRAL

Se propone aplicar la inferencia clásica y la inferencia basada en modelos lineales generalizados marginales y mixtos, para estimar la proporción de alumnos universitarios con vocación emprendedora en la población objetivo. Ella está formada por quienes 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. Teniendo 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[20]. La muestra de alumnos (elemento de muestreo) se ha obtenido seleccionando al azar los cursos (unidad de muestreo) en los cuales efectuar las encuestas.

Ésta 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. La selección de cursos no fue considerada como una etapa adicional del diseño muestral. Por consiguiente, el diseño consta de dos etapas:

- La selección al azar de facultades o *clusters*.
- El relevamiento de todos los alumnos dentro de los cursos seleccionados, que puede considerarse equivalente a la selección al azar de alumnos dentro de cada *cluster*.

En la Tabla 2 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 que se presentan en las secciones subsiguientes.

[20] A excepción del estrato definido por ingeniería en universidades privadas de la Provincia de Buenos Aires, debido a que el marco muestral no contiene ninguna institución de tales características.

TABLA 2: Características 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

La presencia de vocación emprendedora (VE) surge en forma objetiva de las encuestas realizadas. Se define que un individuo tiene vocación emprendedora si alguna vez creó una empresa o posee al momento del relevamiento un proyecto concreto para crearla, resultando la variable de interés de naturaleza binaria[21].

Se encuestaron 948 alumnos avanzados de economía, administración e ingeniería en 9 universidades y 14 facultades de la Ciudad Autónoma de Buenos Aires y de la Provincia de Buenos Aires, durante el segundo cuatrimestre de 2002. De acuerdo a la definición de variables e indicadores, existen tres posibilidades: *(a)* que el alumno haya creado alguna vez una empresa; *(b)* que el alumno posea una idea concreta de negocios sin haber creado su propia empresa; *(c)* que el alumno opine que al graduarse le gustaría crear una empresa, pero al momento no posea ninguna idea de negocios ni haya creado una empresa.

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 ni al grupo objetivo (con VE) ni al grupo control (sin VE). De este modo, la muestra queda compuesta por 799 alumnos: 280 con vocación emprendedora y 519 sin vocación emprendedora (Figura 2).

[21] El formulario de encuesta se presenta en el Anexo A.

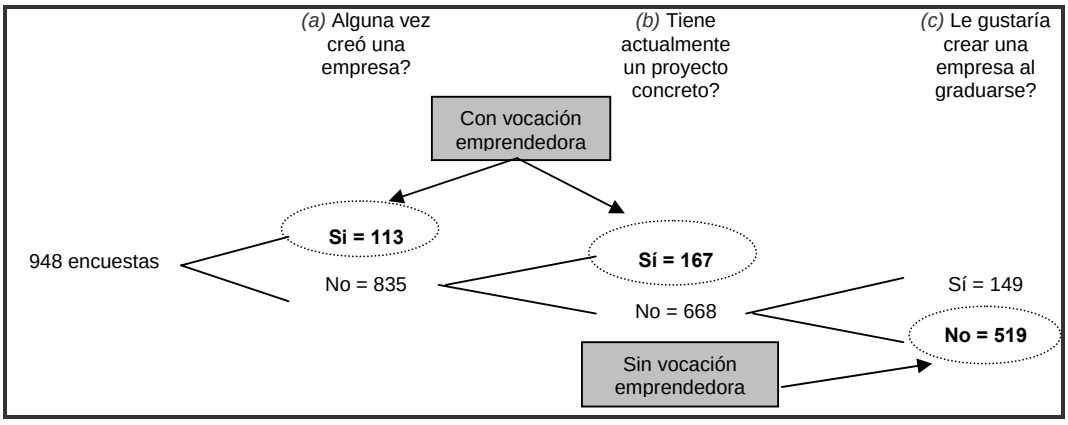

FIGURA 2: Composición de la muestra y definición de la variable respuesta

Del total de 799 encuestas, 723 son procesadas luego de excluir aquéllas con datos faltantes. Su distribución según el tipo de gestión de la facultad y carrera es la siguiente:

- El 86% asiste a facultades de gestión pública y el 14% restante a facultades de gestión privada.
- El 56% cursa carreras de economía y administración, mientras que el otro 44% cursa ingeniería. La distribución por carrera se mantiene por tipo de gestión de la facultad.

6. INFERENCIA BASADA EN DISEÑO MUESTRAL

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 las mismas.

El estimador puntual de la media global $\left(\hat{\mu}_{..}\right)$ es **0.40**, valor que representa la proporción de alumnos universitarios con VE en la población objetivo –i.e., estudiantes avanzados de economía, administración e ingeniería de la Ciudad Autónoma de Buenos Aires y de la Provincia de Buenos Aires–. Entre las medias estimadas por facultad $\left(\hat{\mu}_{i.}\right)$ se observa una alta variabilidad (Tabla 3).

TABLA 3: Proporción estimada de alumnos con vocación emprendedora por facultad

Facultad	Media
U1	0.42
U2	0.24
U3	0.90
U4	0.71
U5	0.19
U6	0.30
U7	0.53
U8	0.50
U9	0.29
U10	0.45
U11	0.50
U12	0.67
U13	0.34
U14	0.26

Aplicando la fórmula [*13*], la varianza de la media global se estima en **0.0007**. Ésta puede descomponerse en dos términos:

- Un primer término, proporcional a la variabilidad entre la media global y la media de *cluster,* de **0.0005**.
- Un segundo término, proporcional a la variabilidad intra-*cluster,* de **0.0002**.

El valor del primer término, 2.9 veces superior al segundo, sugiere que existe una mayor variabilidad entre *clusters*, i.e., son más homogéneas entre sí las respuestas de alumnos de una misma facultad, que las de alumnos de distintas facultades. Los valores estimados se resumen en la Tabla 4.

TABLA 4: Proporción estimada de alumnos con vocación emprendedora

Media global	Varianza	Error estándar	Intervalo de confianza		Amplitud IC	Tamaño muestral
			Lím.inferior	Lím.superior		
0.399	0.0007	0.026	0.348	0.452	0.104	723

Dada la importancia que revisten las variables tipo de gestión de la facultad (X_1) –pública o privada– y carrera cursada (X_2) –economía y administración o ingeniería– dentro del diseño efectuado, es interesante estimar la proporción de alumnos con VE (Y) para los distintos niveles de las mismas. Una forma simple de determinar si dichas proporciones difieren significativamente entre sí, es mediante el uso de tablas de contingencia. En ellas se calculan las probabilidades condicionales de Y para los distintos valores fijos de X, haciendo un análisis marginal sobre las restantes variables.

Lo que se desea analizar es si las distribuciones son o no homogéneas entre sí. Con tal fin puede realizarse una prueba chi-cuadrado bajo la hipótesis nula que establece que las proporciones de alumnos con vocación emprendedora no difieren entre los distintos niveles de la variable de clasificación. Sin embargo, estas pruebas ignoran completamente el diseño muestral, ya que implícitamente suponen que los datos han sido captados mediante un muestreo aleatorio irrestricto (Brier, 1980; Hidiroglou & Rao, 1987a y 1987b). Por lo tanto, resultan poco adecuadas para este caso de estudio.

Realizando un análisis por tipo de gestión de la facultad, en la Tabla 5 se incluyen las medias y varianzas estimadas para cada uno de los dominios, observándose una mayor vocación emprendedora entre los alumnos de instituciones privadas.

TABLA 5: Proporción estimada de alumnos con vocación emprendedora según la gestión pública o privada de la facultad

	Media global	Varianza	Error estándar	Intervalo de confianza		Amplitud IC	Tamaño muestral
				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

A fin de establecer si la diferencia entre las proporciones estimadas resulta estadísticamente significativa, se aplica una prueba de igualdad de proporciones. Dado que se rechaza la hipótesis nula de igualdad, se concluye que la media de alumnos con vocación emprendedora difiere entre las facultades privadas y públicas.

$$z = \frac{0.374 - 0.674}{\sqrt{0.001 + 0.004}} = -4.24$$

.

$$\Pr\left(|z| > 4.24\right) < 0.001$$

La media global y la varianza estimadas por carrera se presentan en la Tabla 6. Como puede apreciarse, la proporción de alumnos con vocación emprendedora es levemente superior en carreras de ciencias económicas que en carreras ingenieriles.

TABLA 6: Proporción estimada de alumnos con vocación emprendedora por carrera

	Media global	Varianza	Error estándar	Intervalo de confianza		Amplitud IC	Tamaño muestral
				Lím.inferior	Lím.superior		
Economía y Administración	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

Mediante la realización de la prueba de igualdad de proporciones, se obtiene que la diferencia hallada no resulta estadísticamente significativa. Por lo tanto, se concluye que las proporciones no difieren por carrera.

$$z = \frac{0.405 - 0.373}{\sqrt{0.001 + 0.002}} = 0.58$$

.

$$\Pr\left(|z| > 0.58\right) = 0.562$$

Asimismo, es de interés calcular las proporciones de alumnos con vocación emprendedora por género. Acá surge una primera cuestión: las fórmulas que se emplean para estimar la media y la varianza suponen el conocimiento del tamaño del *cluster*. En los casos anteriores, esto requería saber cuántos alumnos cursan el último año de las carreras analizadas, lo que puede establecerse de manera aproximada.

El problema de determinar el número exacto de alumnos en esta situación se origina porque durante un cuatrimestre se cursan varias materias y el número de inscriptos en ellas es disímil. Dos criterios pueden utilizarse: *(a)* promediar el número de inscriptos a las distintas materias; *(b)* optar por el número de inscriptos en la materia que los alumnos cursaban al momento de efectuar las encuestas[22].

Cuando se desea estimar la proporción de alumnos con VE por género, el problema es más grave puesto que no existe información disponible acerca de la cantidad de hombres y mujeres que cursan las carreras bajo análisis. Por tal motivo, en este trabajo 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.

Así, en la Tabla 7 se indican la media global y la varianza correspondientes a cada género, calculadas con este supuesto.

TABLA 7: Proporción estimada de alumnos con vocación emprendedora por género

	Media global	Varianza	Error estándar	Intervalo de confianza		Amplitud IC	Tamaño muestral
				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

La realización de la prueba de igualdad de proporciones revela que las diferencias entre las proporciones de alumnos con vocación emprendedora difieren por género.

[22] En muchos casos, las instituciones brindan dicha información en forma aproximada porque el dato reviste carácter estratégico o por la falta de disposición a procesar la información. Estas deficiencias, aunque se presumen, no pueden comprobarse.

$$z = \frac{0.316 - 0.469}{\sqrt{0.001 + 0.001}} = -3.42$$

.

$$\Pr\left(|z| > 3.42\right) < 0.001$$

Para concluir, vale destacar algunas limitaciones que surgen de la aplicación de la inferencia clásica para estimar la proporción de alumnos con vocación emprendedora:

- 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.
- 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 4 a 7, a menor tamaño de muestra, más imprecisa es la estimación.
- En este apartado se ha contemplado una única variable de clasificación: tipo de gestión de la facultad, carrera o género. Si se quisiera utilizar dos o más variables al mismo tiempo, se observaría que el tamaño muestral disminuye notablemente con la estratificación efectuada a posteriori, con el consecuente incremento en la amplitud de los intervalos de confianza.

7. INFERENCIA BASADA EN MODELOS

Uno de los objetivos que se persigue al ajustar un modelo a un conjunto de datos es conocer el proceso generador de los mismos, aunque modelo y proceso no son sinónimos. El modelo representa una aproximación válida del proceso que, mediante un número pequeño de parámetros, permite responder los interrogantes planteados por los datos (Diggle *et al.*, 2002). El proceso de ajuste de un modelo puede dividirse en cuatro etapas: Formulación, Estimación, Inferencia y Diagnóstico.

La formulación consiste en elegir la forma general del modelo. Esencialmente, esta fase continúa el análisis exploratorio, aunque se direcciona hacia aspectos concretos de los datos que el modelo intenta describir –e.g., covariables a incluir, relación matemática o probabilística entre ellas y la variable respuesta–. En esta instancia, el foco de atención lo constituyen la estructura de medias y de dependencia.

Generalmente, el investigador tiene una base teórica que justifica ciertas especificaciones y, en la mayoría de los campos de estudio, existen convenciones acerca de qué covariables deben incluirse en el modelo. Al respecto, Agresti (2002) menciona que la significatividad estadística no debe ser el único criterio para juzgar la inclusión de una variable en el predictor lineal: puede ser sensato incluirla aunque su efecto estimado esté asociado con un alto valor *p*, si ella desempeña un rol central a los fines del estudio. En tal caso, conservarla puede contribuir a reducir el sesgo en los efectos estimados de las otras covariables.

En este trabajo se formulan cuatro modelos lineales generalizados, todos con enlace logístico: *(a)* marginal con distintas estructuras de dependencia; *(b)* mixto con verosimilitud completa; *(c)* mixto con verosimilitud condicional; *(d)* con observaciones independientes. Todos ellos incluyen los mismos efectos fijos de covariables en el predictor lineal.

En la fase de estimación se le asignan valores numéricos a los parámetros del modelo. De acuerdo a la naturaleza del cuerpo de datos y al modelo elegido, se optará por el método de estimación que resulte más adecuado. En este caso, respectivamente, se emplean los

métodos de: *(a)* ecuaciones de estimación generalizadas –basado en el enfoque de cuasi-verosimilitud–; *(b)* verosimilitud completa; *(c)* verosimilitud condicional; *(d)* máxima-verosimilitud.

En la etapa de inferencia se calculan los errores estándares de los parámetros estimados para construir intervalos de confianza y contrastar hipótesis respecto de los mismos. La magnitud de dichos errores y la amplitud de los intervalos de confianza, permiten evaluar la eficiencia de los estimadores. Éste constituye un criterio útil, aunque no el único, para valorar las ventajas y desventajas de las distintas estrategias de modelación.

Por último, durante la fase de diagnóstico se corrobora que el modelo verdaderamente se ajuste a los datos. El objetivo consiste, entonces, en comparar los valores observados con los ajustados a fin de detectar discrepancias significativas –análisis posdictivo–. El poder predictivo del modelo es obtenido para realizar sugerencias sobre su uso futuro.

7.1. Formulación

Definición de covariables

La variable respuesta en este estudio es la presencia de vocación emprendedora en alumnos universitarios que cursan el último año de carreras de economía, administración e ingeniería en la Ciudad Autónoma de Buenos Aires y Provincia de Buenos Aires (República Argentina). Se define que un individuo posee vocación emprendedora si alguna vez inició una empresa propia o si tiene actualmente un proyecto concreto para crearla. Dicha variable es de naturaleza binaria y surge objetivamente de las encuestas realizadas.

La presencia de vocación emprendedora es descripta como función de efectos fijos de covariables y también de efectos aleatorios para algunas de las formulaciones elegidas. Para los efectos fijos el espacio de inferencia está limitado a los niveles observados de dicho efecto, mientras que para los aleatorios el espacio de inferencia se aplica a la población de niveles, no todos los cuales se observan en los datos (Moisen *et al.*, 1999). En este estudio, el

efecto de la facultad a la que concurre el alumno es el que se considera aleatorio bajo los modelos mixtos.

Las covariables contenidas en el predictor lineal, de naturaleza binaria, destacan elementos de distinta naturaleza[23]. El género es algo propio del individuo, la situación laboral representa un factor de tipo situacional, la actitud frente al desempleo y la visión empresarial son elementos subjetivos y, por último, la propensión al riesgo y la creatividad son características en buena parte innatas. La educación universitaria puede ejercer influencia sobre varios de estos factores, se halle o no orientada específicamente hacia el desarrollo de capacidades emprendedoras. En la Tabla 8 se definen las covariables, justificando su inclusión en el modelo desde la teoría de creación de firmas.

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, siendo 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 justifican por qué es válido considerar que las observaciones no son independientes. Por ello, en este trabajo se emplean métodos de análisis que contemplan la dependencia existente entre las mediciones en una misma facultad.

[23] En el Anexo A se presentan todas las variables que surgen de la encuesta.

TABLA 8: Definición de covariables

Rótulo	Descripción y codificación	Hipótesis de trabajo y justificación
GENERO	Indica si el alumno es hombre (1) o mujer (0).	*Los hombres tienen mayores chances de poseer vocación emprendedora.* Estudios empíricos en la temática de creación de firmas concluyen que entre los emprendedores prevalecen los hombres (Reynolds *et al.*, 2002).
OCUPADO	Indica si el alumno está actualmente trabajando (1) o no (0).	*Los individuos ocupados tienen mayores chances de poseer vocación emprendedora.* La experiencia ocupacional, y particularmente aquélla obtenida en pymes o empresas familiares, puede generar interés en la iniciación de una actividad empresarial. Además, el aprendizaje en el puesto de trabajo puede actuar como incubadora de futuros emprendedores, puesto que allí el individuo acumula información a partir de la cual cobra cuerpo su idea (Côté, 1991).
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.* Uno de los incentivos para que los alumnos universitarios se involucren en actividades empresariales se vincula con las condiciones que ofrece el mercado laboral y los diferenciales de ingreso. Si ellos evalúan que el mercado no ofrece empleos acordes a su formación universitaria y que desempeñándose por cuenta propia obtendrán un retorno superior al esperado bajo relación de dependencia, se verán inducidos a pensar en la creación de una empresa propia como opción de carrera (Henrekson & Rosenberg, 2001).
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.* Es de esperar que quienes valoran favorablemente la actividad empresarial como opción de carrera sean más propensos a la creación de una empresa propia.
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.* En este aspecto, es la valoración subjetiva del alumno la que se toma en cuenta. Puesto que es el propio individuo quien decide enfrentar el desafío de crear una nueva firma, cabe esperar que aquéllos menos adversos al riesgo sean más propensos a iniciar una empresa propia (Hisrich, 1988).
CREATIV	Indica si el alumno desarrolla en su tiempo libre alguna actividad creativa (1) –creatividad alta – o no (0) –creatividad media/baja–.	*Los alumnos con un alto nivel de creatividad tienen mayores chances de poseer vocación emprendedora.* Existe evidencia de que los alumnos que poseen destreza manual o *hobbies* técnicos son más propensos a poseer una idea de negocios (Scott & Twomey, 1988). Relacionado con tales aspectos se encuentra la realización de actividades creativas, las cuales implican la posesión de habilidades que pueden contribuir a la vocación emprendedora.

Análisis preliminar de las covariables

Excluyendo los casos con datos faltantes, resulta un total de 723 observaciones a procesar[24]. En la Tabla 9 se resume el porcentaje de respuestas con valor 1 para cada una de las covariables en la totalidad de la muestra –i.e., ignorando la estructura de *clusters*–.

TABLA 9: Porcentaje de encuestas con valor 1 en cada variable

Variable	%
VE 1 = Sí; 0 = No	36.10%
GENERO 1 = Hombre; 0 = Mujer	65.70%
OCUPADO 1 = Ocupado; 0 = Desocupado/Inactivo	54.22%
ACTITUD 1 = Empresarial; 0 = Otra	39.42%
VISIÓN 1 = Favorable; 0 = Desfavorable	75.93%
RIESGO 1 = Propenso; 0 = Adverso	29.46%
CREATIV 1 = Alta; 0 = Media/Baja	27.11%

Antes de ajustar un modelo, es interesante estudiar el efecto que ejerce cada una de las covariables sobre la variable respuesta. Con ese fin, en la Tabla 10 se indican los cocientes de chances marginales estimados –i.e., ignorando el efecto de las restantes covariables– entre vocación emprendedora y cada covariable, junto con sus intervalos de confianza exactos al 95%.

TABLA 10: Cocientes de chances marginales entre la variable respuesta (VE) y las covariables

Covariable	Cociente de chances	Límites de los Intervalos de confianza exactos al 95%	
		Inferior	Superior
GENERO	2.16	1.52	3.09
OCUPADO	2.55	1.83	3.56
ACTITUD	3.84	2.75	5.35
VISION	7.39	4.31	13.34
RIESGO	2.93	2.07	4.13
CREATIV	1.44	1.01	2.04

[24] El coeficiente de correlación calculado entre la variable respuesta y una variable indicadora que le asigna el valor 1 a las observaciones con datos incompletos, es de –0.068. Ello indicaría que los datos faltantes no obedecen a un patrón regular que deba modelarse (Agresti, 2002). El cociente de chances calculado entre VE y la variable indicadora de 0.59, resulta no significativo.

Tal como se desprende de la tabla anterior, la totalidad de las covariables seleccionadas exhiben asociación marginal significativa con la vocación emprendedora –i.e., ningún intervalo de confianza incluye al 1–. De este modo, ignorando la influencia de las restantes covariables, se puede concluir que:

> Las chances de poseer vocación emprendedora son mayores si el alumno es hombre, se halla trabajando actualmente, posee actitud emprendedora frente al desempleo, visualiza favorablemente la actividad empresarial, es propenso al riesgo o tiene alta creatividad.

La relación entre las covariables y la variable respuesta puede analizarse mediante las pruebas de asociación homogénea de Breslow-Day (B-D) y de independencia condicional de Cochran-Mantel-Haenszel (CMH), usando tablas de contingencia a múltiples vías de clasificación[25]. Cuando la asociación se mantiene estable en las diversas tablas parciales, la información puede combinarse en el estimador de Mantel-Haenszel (MH) de un cociente de chances común como medida resumen de asociación condicional:

$$MH = \frac{\sum_{i=1}^{k} \frac{n_{11i} n_{22i}}{n_{++i}}}{\sum_{i=1}^{k} \frac{n_{12i} n_{21i}}{n_{++i}}} \quad . \qquad [75]$$

Asimismo, SAS PROC FREQ ofrece un estimador LOGIT que usa una corrección de 0.5 en cada celda de las tablas parciales que contienen un cero, aunque excluye del cómputo del cociente de chances común a aquellas tablas con un cero en el total de la fila o de la columna. Los estimadores MH y LOGIT, no basados en el modelo, ofrecen ventajas respecto de la estimación basada en el modelo, $exp\left(\hat{\beta}\right)$, cuando el número de tablas parciales es elevado y éstas contienen pocos datos.

[25] Las hipótesis nulas son, respectivamente, “Existe asociación homogénea” y “Existe independencia condicional”.

En la Tabla 11 se presentan los valores *p* asociados a las referidas pruebas y los cocientes de chances común estimados. Para ello se utilizan las 2^5 = 32 tablas parciales que resultan de cruzar la totalidad de las covariables binarias, ignorando la estructura de *clusters*.

TABLA 11: Prueba de asociación homogénea, prueba de independencia condicional y cociente de chances común ignorando el efecto de *cluster*

Tabla de contingencia parcial	Valor *p* de B-D	Valor *p* de CMH	Cociente de chances común*	
			MH	LOGIT
GENERO * VE	0.059	<0.001	2.17	2.00
OCUPADO * VE	0.008	<0.001	2.78	2.71
ACTITUD * VE	0.074	<0.001	2.81	2.77
VISION * VE	0.443	<0.001	4.73	3.48
RIESGO * VE	0.004	<0.001	2.27	2.33
CREATIV * VE	0.024	0.011	1.63	1.63

*Todos los cocientes de chances son estadísticamente significativos.

Tal como surge de la tabla anterior:

- En ninguno de los casos se cumple el supuesto de independencia condicional (se rechaza al prueba de CMH con α = 0.05).
- Existe asociación homogénea entre GENERO, ACTITUD y VISION con VE, controlando por las restantes covariables (no se rechaza la prueba de B-D con α = 0.05).
- La asociación entre OCUPADO, RIESGO y CREATIV con VE no se mantiene constante en los distintos niveles de las demás covariables (se rechaza la prueba de B-D con α = 0.05).

No obstante, tal como explica Agresti (2002), aún con un bajo valor *p* en la prueba de asociación homogénea, si la variabilidad en los cocientes de chances estimados no es sustancial, una medida resumen como el cociente de chances común sigue siendo útil. El cumplimiento del supuesto de asociación homogénea no es un prerrequisito para calcular dicha medida ni para probar la existencia de independencia condicional.

En la Tabla 12 se repite el análisis anterior, contemplando ahora el efecto de *cluster* al considerarlo como una vía más de clasificación, obteniéndose así 32x14 = 448 tablas parciales en cada caso.

TABLA 12: Prueba de asociación homogénea, prueba de independencia condicional y cociente de chances común incorporando el efecto de *cluster*

Tabla de contingencia parcial	Valor *p* de B-D	Valor *p* de CMH	Cociente de chances común	
			MH	LOGIT
GENERO * VE	0.828	<0.001	7.30	3.29
OCUPADO * VE	0.015	<0.001	2.87	2.28
ACTITUD * VE	0.012	<0.001	2.82	2.47
VISION * VE	0.270	<0.001	5.12	2.57
RIESGO * VE	0.059	<0.001	2.30	2.15
CREATIV * VE	0.002	0.053	1.55*	1.48*

*El cociente de chances común estimado resulta estadísticamente significativo con α=0.10.

Teniendo en cuenta los *clusters* como una vía más de clasificación, resulta que:

- Sigue sin cumplirse el supuesto de independencia condicional (se rechaza la prueba de CMH con α = 0.10).
- Continúa habiendo asociación homogénea entre GENERO y VISION con VE, controlando por las restantes covariables, y ahora este supuesto se cumple también para RIESGO (no se rechaza la prueba de B-D con α = 0.05).
- La asociación entre OCUPADO, ACTITUD y CREATIV con VE, teniendo en cuenta las demás covariables, no se mantiene constante en las distintas tablas parciales (se rechaza la prueba de B-D con α = 0.05).

Los cocientes de chances común MH estimados, a excepción del correspondiente a GENERO –covariable que presenta la mayor variabilidad entre los *clusters*–, son aproximadamente iguales a los anteriores. En cuanto a los estimadores LOGIT, si se compara los valores que asumen en la Tabla 11 con los de la Tabla 12, se observa que disminuyen en mayor medida y difieren más de los estimadores MH. La causa que puede explicar las diferencias entre ambos es que, trabajando con tablas de grandes dimensiones, hay una gran cantidad de tablas con ceros en el total de la fila o de la columna que son excluidas del cómputo del estimador LOGIT.

En principio, podría suponerse que la falta de cumplimiento del supuesto de asociación homogénea invalida el uso de un modelo de efectos principales y que debería contemplarse la presencia de interacciones en el predictor lineal. Sin embargo, ya se ha mencionado que dicha condición no es fundamental para estimar un cociente de chances común. Por

consiguiente, inicialmente se formulará un modelo sin interacciones y luego se evaluará si la adición de las mismas mejora significativamente el ajuste.

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*:

$$g\left(\mu_{ij}\right) = \text{logit}\left(\mu_{ij}\right) = \eta_{ij}$$
$$i = 1,...,k \quad ,$$
$$j = 1,...,m_i$$

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

$$\text{corr}\left(y_{ir}, y_{is}\right) = \rho\left(\mu_{ir}, \mu_{is}, \boldsymbol{\alpha}\right) \qquad \forall r \neq s\,.$$

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*:

$$g\left(\mu_{ij} / U_i\right) = \text{logit}\left(\mu_{ij} / U_i\right) = \eta_{ij}$$
$$i = 1,...,k \quad ,$$
$$j = 1,...,m_i$$

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

$$U_i \sim N\left(0, \sigma_u^2\right).$$

Bajo el **modelo mixto con verosimilitud condicional**, las covariables deben asumir distintos valores dentro de cada *cluster* y, a diferencia del modelo anterior, no es preciso especificar una distribución paramétrica para los efectos aleatorios. La función de enlace también es la logística:

$$g\left(\mu_{ij}/U_i\right) = \text{logit}\left(\mu_{ij}/U_i\right) = \eta_{ij}$$
$$i = 1,\ldots,k \quad ,$$
$$j = 1,\ldots,m_i$$

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

Por último, el **modelo de regresión logística ordinario**, supone independencia entre las observaciones e ignora completamente la estructura de *clusters*:

$$g\left(\mu_{ij}\right) = \text{logit}\left(\mu_{ij}\right) = \eta_{ij}$$
$$i = 1,\ldots,k \quad ,$$
$$j = 1,\ldots,m_i$$

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

7.2. Estimación

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, NLMIXED y PHREG. La definición de CLUSTER y SUBCLUSTER es la siguiente:

- **CLUSTER**: los individuos se agrupan de acuerdo a la facultad a la que asisten. Esta variable posee 14 niveles.
- **SUBCLUSTER**: los individuos se agrupan teniendo en cuenta la facultad y la titulación, uniéndose las titulaciones que resultan afines entre sí –e.g., ingeniería electrónica e ingeniería electromecánica–, con el fin de evitar unidades de tamaño muy pequeño. Esta variable posee 25 niveles[26].

El tamaño de muestra por *cluster* varía de 6 a 137 con una media de 52 alumnos y un desvío estándar de 42 (CV=81%), mientras que por *subcluster* varía de 3 a 92 con una media de 29 alumnos y un desvío estándar de 23 (CV=79%)[27]. La variabilidad presente en la muestra es un reflejo de la que existe a nivel poblacional, puesto que las facultades difieren ampliamente en cuanto al número de alumnos.

[26] En la Tabla A-3 del Anexo A se detalla la conformación de los *subclusters*.
[27] CV representa al coeficiente de variación, igual al cociente entre el desvío estándar y la media.

Con variables explicativas categóricas, existen tantos parámetros asociados a cada una de ellas como niveles o categorías, uno de los cuales es redundante. En este caso, sólo se reporta el estimador $\left(\hat{\beta}\right)$ asociado al valor 1 de la covariable y, dada la parametrización adoptada, $exp\left(\hat{\beta}\right)$ representa el cociente entre las chances de VE=1 cuando *X=1* y las chances de VE=1 cuando *X=0*, para cada covariable.

Análisis de multicolinealidad

Un modelo con más de una covariable puede presentar multicolinealidad. Si el coeficiente de correlación entre los predictores es alto, es posible que alguna covariable incorrectamente parezca no ejercer influencia cuando las demás están en el modelo, debido a que los efectos de ésta se superponen con los de otra u otras covariables (Agresti, 1996). Como consecuencia, los errores estándares de los estimadores son mayores y los parámetros se estiman ineficientemente.

A fin de analizar esta cuestión, se reproduce en la Tabla 13 la matriz de correlación entre las covariables –realizada con SAS PROC CORR–, en la que puede observarse que todos los coeficientes, en valor absoluto, son inferiores a 0.25.

TABLA 13: Matriz de correlación

	GENERO	OCUPADO	ACTITUD	VISION	RIESGO	CREATIV
GENERO	1.00	-0.09*	0.08*	0.13*	0.10*	-0.11*
OCUPADO		1.00	0.05	0.10*	0.04	-0.06
ACTITUD			1.00	0.23*	0.18*	0.04
VISION				1.00	0.13*	0.05
RIESGO					1.00	0.05
CREATIV						1.00

* Indica que el coeficiente de correlación estimado es significativo (α = 0.05).

Cabe aclarar que con más de dos covariables la medición de la multicolinealidad se torna complicada, ya que puede ser alta aún si la correlación de a pares no es elevada (Kmenta, 1977). Por tal motivo, en la Tabla 14 se presentan los índices de condición –raíz cuadrada del cociente entre el máximo autovalor y cada autovalor correspondiente a la matriz *(X'X)*–, como

medida de multicolinealidad. Ninguno de los índices supera el valor de 30 –a partir del cual se sospecha que la misma es de moderada a alta–, descartándose así la existencia de multicolinealidad entre las covariables seleccionadas para la modelización.

TABLA 14: Índices de condición

Dimensión	Autovalores	Índices de condición
1	4.38	1.00
2	0.73	2.45
3	0.63	2.64
4	0.51	2.92
5	0.44	3.17
6	0.22	4.46
7	0.10	6.56

En consonancia con el análisis realizado en el apartado anterior, resulta interesante evaluar la asociación entre las covariables mediante los valores que asumen los cocientes de chances marginales entre pares (Tabla 15). Puede observarse que la asociación entre las covariables tiende a ser débil y aún no significativa en varios casos –cuando el intervalo contiene al 1–. La mayor asociación se evidencia entre ACTITUD y VISION, lo que indica que quienes poseen una visión favorable de la actividad empresarial tienen más chances de tener una actitud emprendedora frente a una situación de desempleo (o viceversa).

TABLA 15: Cocientes de chances marginales entre covariables

	OCUPADO	ACTITUD	VISION	RIESGO	CREATIV
GENERO	0.70*	1.39*	1.81*	1.64*	0.60*
OCUPADO		1.22	1.59*	1.19	0.77
ACTITUD			3.47*	2.22*	1.21
VISION				2.08*	1.28
RIESGO					1.27

* Indica que el cociente de chances marginal estimado es significativo ($\alpha = 0.05$).

Modelo marginal

Bajo el enfoque marginal, el modelo se estimó en SAS PROC GENMOD con el método de ecuaciones de estimación generalizadas *(GEE)*, considerando distintas alternativas para modelar la falta de independencia entre las observaciones. El cociente entre el estadístico X^2

de Pearson, igual a 734.87, y sus 716 grados de libertad, da como resultado 1.03, lo que permite descartar la presencia de sobredispersión.

De las estimaciones efectuadas, primero se exponen las obtenidas al modelar la matriz de correlación de trabajo con las estructuras de independencia (TYPE=IND) y de simetría compuesta (TYPE=EXCH). No se ha utilizado la matriz sin estructura (TYPE=UN) debido a que el número de pares dentro de alguno de los *clusters* es menor que el número de parámetros a estimar, por lo que se ha optado por modelos más simples.

Los valores iniciales de los parámetros, a partir de los cuales se obtienen las estimaciones por *GEE*, suponen que las observaciones no se hallan correlacionadas y, por consiguiente, coinciden con las contenidas en la Tabla 16. A pesar de que la estructura de correlación de trabajo sea la de independencia (TYPE=IND), los errores estándares difieren de los valores iniciales. Ello se deba a que en *GEE* la estructura propuesta en la matriz de correlación de trabajo no representa, necesariamente, la estructura de correlación de las observaciones –ver [*38*]–. Las pruebas de tipo III realizadas indican que el cambio en el ajuste, si cualquiera de las covariables se excluye del modelo, es significativo en presencia de las demás covariables –con un valor *p* de 0.10 para CREATIV–[28].

TABLA 16: Estimación del modelo marginal con estructura de correlación de independencia –TYPE=IND–

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* prueba de cuasi-score*
INTERCEPTO	–3.919	0.442	<0.001	0.02	
GENERO	0.801	0.181	<0.001	2.23	0.027
OCUPADO	1.056	0.243	<0.001	2.87	0.022
ACTITUD	1.074	0.128	<0.001	2.93	0.006
VISION	1.602	0.329	<0.001	4.96	0.009
RIESGO	0.818	0.172	<0.001	2.26	0.006
CREATIV	0.484	0.232	0.037	1.62	0.100

* Probabilidad asociada a los estadísticos de cuasi-*score* para pruebas de tipo III.

Ajustando el modelo con la estructura de correlación de simetría compuesta (TYPE=EXCH), la correlación común estimada entre pares de respuestas es baja (0.025). Esto implica que las observaciones de a pares se comportan como si fueran independientes y, en consecuencia, las estimaciones no difieren significativamente de las obtenidas con una regresión logística ordinaria. No obstante, los errores estándares calculados por este método

están incorporando la dependencia empírica entre las respuestas y son, por ello, más robustos[29]. Los valores estimados se presentan en la Tabla 17.

TABLA 17: Estimación del modelo marginal con estructura de correlación de simetría compuesta –TYPE=EXCH–

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* prueba de cuasi-score*
INTERCEPTO	–3.804	0.363	<0.001	0.02	
GENERO	0.930	0.152	<0.001	2.53	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

* Probabilidad asociada a los estadísticos de cuasi-*score* para pruebas de tipo III.

Todas las covariables son estadísticamente significativas según la prueba de Wald. Este resultado se confirma con la prueba de cuasi-*score* –más conservadora–, la cual indica que el cambio en el ajuste es significativo si alguna covariable se excluye del modelo en presencia de las restantes.

Como segunda estrategia se ha modelado la dependencia mediante cocientes de chances, aplicando el método *ALR*. Se han considerado tres alternativas: simetría compuesta (LOGOR=EXCH), cocientes de chances anidados a un nivel (LOGOR=NEST1) y cocientes de chances por *cluster* (LOGOR=LOGORVAR).

En el primer caso (LOGOR=EXCH) se estima un único cociente de chances que se supone común a todos los pares de observaciones dentro de cada *cluster*. En el segundo caso (LOGOR=NEST1), en el que la variable anidada dentro de los *clusters* son los *subclusters*, se estima: *(i)* un cociente de chances para individuos que comparten *cluster* y *subcluster* –alumnos de la misma facultad y titulación– y *(ii)* otro cociente de chances para individuos del mismo *cluster* pero distinto *subcluster* –alumnos de la misma facultad pero distinta titulación–.

En el tercer caso se estiman cocientes de chances que son constantes al interior del *cluster* pero que asumen distintos valores para cada uno de los niveles de las siguientes variables argumento: ZONA, la cual representa la localización geográfica de la facultad

[28] La hipótesis nula plantea que el cambio en el ajuste no es significativo si la covariable se excluye del modelo.
[29] Al respecto, Agresti (2002) presenta un caso en el cual la correlación obtenida es de –0.003, en la que utiliza este mismo argumento para justificar el uso del modelo marginal.

(zona1: Ciudad Autónoma de Buenos Aires y Gran Buenos Aires; zona 2: resto de la Provincia de Buenos Aires) y CARRERA (economía y administración; ingeniería).

Tal como se observa en la última fila de la Tabla 18, el cociente de chances común para la estructura de simetría compuesta (LOGOR=EXCH) de 1.11 evidencia una débil asociación entre cualquier par de observaciones. No obstante, debe destacarse que el mismo es estadísticamente significativo, lo cual justifica el uso de esta estrategia de modelación que contempla la dependencia entre las respuestas.

TABLA 18: Estimación del modelo marginal con estructura de logaritmos de cocientes de chances intercambiables –LOGOR=EXCH–

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* prueba de cuasi-score*
INTERCEPTO	–3.872	0.380	<0.001	0.02	
GENERO	0.907	0.157	<0.001	2.48	0.026
OCUPADO	1.021	0.233	<0.001	2.77	0.027
ACTITUD	1.050	0.129	<0.001	2.86	0.007
VISION	1.562	0.317	<0.001	4.77	0.009
RIESGO	0.777	0.170	<0.001	2.17	0.007
CREATIV	0.510	0.224	0.023	1.66	0.086
ALFA 1	0.104	0.041	0.011	1.11	

* Probabilidad asociada a los estadísticos de cuasi-*score* para pruebas de tipo III.

Los valores de exp(ALFA 1) y exp(ALFA 2), ambos significativos, representan los dos cocientes de chances que surgen al modelar la asociación mediante cocientes de chances anidados a nivel *subcluster* (LOGOR=NEST1) (Tabla 19). Dado que sus valores prácticamente coinciden, ello implica que las respuestas de alumnos de una misma facultad se encuentran igualmente asociadas, independientemente de la titulación.

TABLA 19: Estimación del modelo marginal con estructura de logaritmos de cocientes de chances anidados a un nivel –LOGOR=NEST1–

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* prueba de cuasi-score*
INTERCEPTO	–3.881	0.384	<0.001	0.02	
GENERO	0.914	0.159	<0.001	2.49	0.027
OCUPADO	1.024	0.232	<0.001	2.78	0.026
ACTITUD	1.049	0.128	<0.001	2.85	0.007
VISION	1.561	0.316	<0.001	4.76	0.009
RIESGO	0.777	0.170	<0.001	2.17	0.007
CREATIV	0.518	0.224	0.021	1.68	0.085
ALFA 1	0.092	0.049	0.062	1.10	
ALFA 2	0.122	0.042	0.003	1.13	

* Probabilidad asociada a los estadísticos de cuasi-*score* para pruebas de tipo III.

Utilizando como estructura de asociación los logaritmos de cocientes de chances por bloque (LOGOR=LOGORVAR) para la variable ZONA, se obtienen dos cocientes de chances constantes al interior de los *clusters*: exp(ALFA 1) correspondiente a la zona 1 y exp(ALFA 2) correspondiente a la zona 2 (Tabla 20). Sólo el primero de ellos resulta significativo, a partir de lo cual se puede inferir que existe asociación entre las observaciones de la Ciudad Autónoma de Buenos Aires y del Gran Buenos Aires, pero no entre las pertenecientes al resto de la Provincia.

TABLA 20: Estimación del modelo marginal con estructura de logaritmos de cocientes de chances por zona –LOGOR=LOGORVAR(ZONA)–

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* prueba de cuasi-score*
INTERCEPTO	–4.092	0.408	<0.001	0.02	
GENERO	0.913	0.143	<0.001	2.49	0.026
OCUPADO	1.064	0.240	<0.001	2.90	0.027
ACTITUD	1.062	0.133	<0.001	2.89	0.007
VISION	1.591	0.364	<0.001	4.91	.
RIESGO	0.765	0.176	<0.001	2.15	0.008
CREATIV	0.531	0.216	0.014	1.70	0.072
ALFA 1	0.2266	0.0501	<0.001	1.25	
ALFA 2	–0.0210	0.0232	0.365	0.97	

* Probabilidad asociada a los estadísticos de cuasi-*score* para pruebas de tipo III.

Si la variable seleccionada como bloque es CARRERA, los dos cocientes de chances estimados son: exp(ALFA 1) para las titulaciones ingenieriles y exp(ALFA 2) para las titulaciones de economía y administración (Tabla 21). El único que resulta estadísticamente significativo es el correspondiente a ALFA 1, por lo que se concluye que hay asociación entre las mediciones de alumnos de ingeniería, pero no así entre quienes cursan carreras de ciencias económicas.

TABLA 21: Estimación del modelo marginal con estructura de logaritmos de cocientes de chances por carrera –LOGOR=LOGORVAR(CARRERA)–

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* prueba de cuasi-score*
INTERCEPTO	–3.889	0.382	<0.001	0.02	
GENERO	0.896	0.158	<0.001	2.45	0.031
OCUPADO	1.020	0.235	<0.001	2.77	0.026
ACTITUD	1.051	0.129	<0.001	2.86	0.007
VISION	1.564	0.323	<0.001	4.78	0.009
RIESGO	0.778	0.170	<0.001	2.18	0.007
CREATIV	0.505	0.225	0.025	1.66	0.094
ALFA 1	0.076	0.077	0.3254	1.08	
ALFA 2	0.119	0.044	0.0070	1.13	

* Probabilidad asociada a los estadísticos de cuasi-*score* para pruebas de tipo III.

Modelando la asociación mediante los logaritmos de los cocientes de chances, las covariables incluidas en el predictor lineal resultan en su totalidad estadísticamente significativas. Asimismo, las pruebas de cuasi-*score* indican que, hallándose las restantes covariables en el modelo, ninguna de ellas debe excluirse del mismo.

También se han efectuado los contrastes “GENERO=0, OCUPADO=0”, “ACTITUD=0, VISION=0” y “RIESGO=0, CREATIV=0”, bajo la hipótesis nula que establece que los parámetros asociados a los efectos incluidos en el contraste son simultáneamente iguales a cero. En todos los casos se rechaza la hipótesis nula, concluyendo que las covariables son significativamente distintas de cero, sea que las pruebas se efectúen con 1 o con 2 grados de libertad.

El contraste que propone que todas las covariables son simultáneamente iguales a cero, con 6 grados de libertad, posee asociado un mayor valor *p*, cualquiera sea la modelación de la estructura de dependencia. Al respecto, Agresti (2002) explica que si se desea contrastar la independencia condicional, las pruebas cuya hipótesis nula establece que $\beta = 0$ con 1 grado de libertad (Tablas 16 a 21) son más potentes que las pruebas de bondad del ajuste, i.e., aquéllas en las que el vector de parámetros es igual a cero.

Modelo mixto con verosimilitud completa

Bajo el enfoque mixto basado en la función de verosimilitud completa, el modelo se estimó en SAS PROC NLMIXED. Se consideró, alternativamente, a las variables CLUSTER y SUBCLUSTER para agrupar a las observaciones, exponiéndose en la Tabla 22 los estadísticos de bondad del ajuste calculados en cada caso.

TABLA 22: Estadísticos de bondad del ajuste correspondientes al modelo mixto con verosimilitud completa (NLMIXED)

Agrupamiento observaciones	Criterio	Valor del estadístico
CLUSTER	–2 LOG L	**741.1**
	AIC	**757.1**
	BIC	**762.2**
SUBCLUSTER	–2 LOG L	746.0
	AIC	762.0
	BIC	773.4

Dado que los valores de estos estadísticos son menores cuando se considera a las observaciones agrupadas a nivel CLUSTER, dicha variable es utilizada como efecto aleatorio en el modelo. Vale mencionar que al intentar utilizar la variable UNIVERSIDAD como efecto aleatorio, se ha presentado una dificultad computacional que impidió finalizar el proceso de ajuste. Ésta persistió a pesar de probar el uso de los distintos algoritmos de optimización –*Quasi-Newton* (QUANEW) y *Conjugate Gradient* (CONGRA)– y de las distintas fórmulas disponibles *(updates)*. Este problema no se manifestó en SAS PROC GENMOD al ajustar un modelo marginal considerando a las observaciones agrupadas al interior de la UNIVERSIDAD.

Para ajustar el modelo mixto con verosimilitud completa, se han utilizado como valores iniciales los coeficientes estimados por el modelo de regresión logística ordinaria. Todas las covariables resultan estadísticamente significativas (Tabla 23).

TABLA 23: Estimación del modelo mixto con verosimilitud completa (NLMIXED)

Covariable	β	Error Estándar	Valor *p* 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	

La varianza de los efectos aleatorios, $SIGMA^2 = 0.31$, representa la heterogeneidad no atribuible a las covariables. Dicha varianza no se estima con buena eficiencia, máxime si hay pocos *clusters*, ya que la realización de una aproximación normal de las varianzas implícita en

la prueba de Wald no es recomendable. Por lo tanto, resulta conveniente estimar SIGMA, quien resulta estadísticamente significativa.

Dado que se está ajustando el modelo de intercepto aleatorio, en la Tabla 24 se indican los valores estimados de los efectos aleatorios y de los interceptos por *cluster*. Aplicando la función de enlace inversa [*24*], el valor incluido en la última columna representa las chances asociadas a la presencia de vocación emprendedora para cada *cluster*, en la subpoblación donde *X=0*.

TABLA 24: Efectos aleatorios e interceptos estimados bajo el modelo mixto con verosimilitud completa (NLMIXED)

Facultad	U_i	β_0+U_i	$g^{-1}(\beta_0+ U_i)$
U1	0.299	–3.639	0.026
U2	–0.315	–4.252	0.014
U3	0.836	–3.101	0.043
U4	0.617	–3.320	0.035
U5	–0.522	–4.459	0.011
U6	–0.178	–4.115	0.016
U7	0.361	–3.576	0.027
U8	0.119	–3.818	0.022
U9	–0.775	–4.712	0.009
U10	–0.176	–4.113	0.016
U11	–0.007	–3.944	0.019
U12	0.383	–3.554	0.028
U13	–0.216	–3.959	0.015
U14	–0.404	–4.341	0.013

Puesto que para el modelo basado en la función de verosimilitud completa es factible calcular medidas tales como AIC y BIC para juzgar el ajuste, en la Tabla 25 se presenta el valor que asumen las mismas si las covariables se excluyen de a una del predictor lineal. Ambos criterios destacan como superior al modelo completo, aún cuando AIC tiende a privilegiar los modelos sobreajustados y BIC favorece el uso de menos covariables. La exclusión de cualquiera de las covariables del modelo, en especial VISION y ACTITUD, incrementa el valor de las medidas.

TABLA 25: Estadísticos de bondad del ajuste correspondientes al modelo mixto con verosimilitud completa (NLMIXED)

Covariable removida	AIC	BIC
MODELO COMPLETO	**757.1**	**762.2**
GENERO	774.6	779.1
OCUPADO	784.9	789.4
ACTITUD	789.9	794.4
VISION	792.6	797.1
RIESGO	771.0	775.5
CREATIV	762.0	766.5

En cuanto a los contrastes simultáneos efectuados "GENERO=0, OCUPADO=0", "ACTITUD=0, VISION=0" y "RIESGO=0, CREATIV=0", el valor *p* obtenido es menor a 0.01 en todos los casos. Bajo este enfoque, se efectúan pruebas *F* de Wald –todas ellas con 2 grados de libertad en el numerador y 13 en el denominador–, bajo la hipótesis nula que establece que ambas covariables son simultáneamente iguales a cero[30]. Vale aclarar que en este caso los contrastes son condicionales, puesto que se refieren a parámetros de un mismo *cluster*.

Modelo mixto con verosimilitud condicional

Bajo el enfoque mixto basado en la función de verosimilitud condicional, el modelo se estimó en SAS PROC PHREG considerando a las observaciones correlacionadas dentro de la variable CLUSTER. Los estadísticos de significancia del modelo permiten rechazar la hipótesis nula que establece que los parámetros son simultáneamente iguales a cero (Tabla 26).

TABLA 26: Pruebas de significancia del modelo correspondientes al modelo mixto con verosimilitud condicional (PHREG)

Prueba	Valor observado	Grados de libertad	Valor *p*
Cociente de verosimilitud	178.3	6	<0.001
Score	156.8	6	<0.001
Wald	120.8	6	<0.001

[30] SAS PROC NLMIXED utiliza el método Delta para aproximar la matriz de covarianzas de las expresiones incluidas en el contraste (SAS Institute Inc., 1999a). Este método se basa en la fórmula de Taylor que aproxima por un polinomio a una función infinitamente diferenciable (Cantor, 1997).

Los parámetros estimados, sus errores estándares y sus correspondientes valores *p*, se presentan en la Tabla 27. Como puede observarse, la totalidad de las covariables resultan estadísticamente significativas.

TABLA 27: Estimación del modelo mixto con verosimilitud condicional (PHREG)

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)
GENERO	1.047	0.237	<0.001	2.85
OCUPADO	1.003	0.196	<0.001	2.73
ACTITUD	1.078	0.187	<0.001	2.94
VISION	1.586	0.297	<0.001	4.88
RIESGO	0.756	0.199	<0.001	2.13
CREATIV	0.572	0.209	0.006	1.77

Debe notarse que al condicionar sobre los efectos aleatorios para estimar los efectos fijos, ya no es posible obtener una estimación del intercepto del modelo. Ello impone restricciones al análisis, puesto que impide, e.g., estimar las probabilidades a partir de los coeficientes de regresión. Por tal motivo, se propone como vía indirecta para hallar un valor aproximado del intercepto, aprovechar la relación comentada en el marco conceptual que existe entre el enfoque con verosimilitud condicional y el análisis de sobrevida.

Es posible pensar en la aparición de vocación emprendedora (VE) como un suceso que se verifica en el tiempo donde, empleando la terminología propia del análisis de sobrevida, VE=1 representa al evento y VE=0 al dato censurado. Cuando el tiempo no se observa en forma continua, sino sólo entre dos seguimientos consecutivos, a los datos se los denomina censurados a intervalos, siendo los tiempos de sobrevida discretos. El modelo que se utiliza en estos casos es el logístico para chances discretas *(discrete hazard)*, el cual difiere mínimamente del modelo logístico usual (Fahrmeir & Tutz, 2001).

Bajo estas condiciones, existe una forma de estimar el intercepto del predictor lineal de una transformación log-log de la probabilidad de reaparición del evento. Dado un individuo para el cual todas las covariables son nulas, a partir del cálculo de la función de sobrevida $S_0(t_s)$ en el tiempo t_s, resulta (Collett, 1994):

$$\beta_0 = \log\left\{-\log S_0\left(t_s\right)\right\}. \qquad [76]$$

Si bien este estimador no es exactamente el que corresponde al intercepto del modelo original, el mismo podría utilizarse como aproximación al valor buscado. Esta propuesta se fundamenta en dos cuestiones:

- Los valores hallados de esta forma son muy próximos a los obtenidos con el enfoque de verosimilitud completa, lo que indica que el método resulta útil.
- Aún cuando el valor distara del verdadero valor para el intercepto –el cual permanece desconocido–, esta aproximación permite calcular las medias individuales, lo que de otro modo queda vedado.

Luego de hallar los valores de $S_0(t_s)$ para cada *cluster* y aplicar la transformación indicada en [*76*], se obtienen los interceptos estimados (Tabla 28). No se observan mayores diferencias al comparar estos valores con los correspondientes al modelo mixto con verosimilitud completa (Tabla 24).

TABLA 28: Valor aproximado del intercepto aleatorio para cada facultad bajo el modelo mixto con verosimilitud condicional (PHREG)

Facultad	$S_0(t_s)$	β_0
U1	0.983	–4.089
U2	0.992	–4.867
U3	0.895	–2.198
U4	0.962	–3.259
U5	0.993	–4.918
U6	0.989	–4.514
U7	0.977	–3.750
U8	0.978	–3.805
U9	0.996	–5.555
U10	0.990	–4.608
U11	0.991	–4.649
U12	0.978	–3.829
U13	0.990	–4.577
U14	0.993	–4.912

Seguidamente se calculan distintas medidas de bondad del ajuste para el modelo completo y con covariables excluidas del predictor lineal (Tabla 29). Tanto AIC como SBC favorecen al modelo completo, puesto que las medidas incrementan su valor si se suprime alguna covariable. Al igual que en el modelo basado en la función de verosimilitud completa, el mayor impacto sobre el ajuste corresponde a la remoción de VISION y ACTITUD.

TABLA 29: Estadísticos de bondad del ajuste correspondientes al modelo mixto con verosimilitud condicional (PHREG)

Covariable removida	AIC	SBC
MODELO COMPLETO	**668.9**	**690.3**
GENERO	687.5	705.3
OCUPADO	694.3	712.1
ACTITUD	701.0	718.9
VISION	702.5	720.3
RIESGO	681.4	699.2
CREATIV	674.4	692.2

Regresión logística ordinaria

Aún resta por ver cómo incide ignorar la correlación entre las observaciones, motivo por el cual a continuación se estiman dos modelos de regresión logística ordinaria: *(i)* uno que sólo incluye a los efectos fijos de covariables en el predictor lineal y *(ii)* otro que adiciona en el predictor lineal a los *clusters* como efectos fijos.

Bajo la primera alternativa, comparable con los resultados hallados aplicando el enfoque marginal, se indica en la Tabla 30 el resultado de la estimación. Tal como se aprecia, los coeficientes hallados coinciden con los obtenidos al utilizar la estructura de independencia (Tabla 16). Ello se debe a que las estimaciones iniciales de los parámetros y sus errores estándares, a partir de los cuales se obtienen los estimadores *GEE*, responden al supuesto de observaciones no correlacionadas y son idénticos a los que se obtienen mediante una regresión logística ordinaria. Sin embargo, bajo esta estrategia, los errores estándares no incorporan la dependencia que surge de la muestra y son por ello menos robustos.

TABLA 30: Estimación del modelo de regresión logística ordinaria

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* de LR*
INTERCEPTO	–3.919	0.352	<0.001	0.02	
GENERO	0.801	0.202	<0.001	2.28	<0.001
OCUPADO	1.056	0.186	<0.001	2.87	<0.001
ACTITUD	1.074	0.181	<0.001	2.93	<0.001
VISION	1.602	0.285	<0.001	4.96	<0.001
RIESGO	0.818	0.191	<0.001	2.27	<0.001
CREATIV	0.484	0.202	0.016	1.62	0.016

* Probabilidad asociada a los estadísticos del cociente de verosimilitud para pruebas de tipo III.

Respecto de la segunda alternativa, al introducir variables dicotómicas en el predictor lineal para tratar a los efectos aleatorios como fijos, es necesario estimar un número sustancial de parámetros adicionales. Esto ocasiona problemas, en particular, cuando los *clusters* son de tamaño pequeño a moderado, siendo el enfoque de efectos aleatorios más parsimonioso al permitir que la estimación se realice aún si existen pocas observaciones en cada *cluster* (Fahrmeir & Tutz, 2001). Los resultados obtenidos bajo esta alternativa se incluyen en la Tabla 31.

TABLA 31: Estimación del modelo de regresión logística con observaciones independientes y *clusters* como efectos fijos

Covariable	β	Error Estándar	Valor *p* de Wald	exp(β)	Valor *p* de LR*
INTERCEPTO	–4.558	0.484	<0.001	0.01	
GENERO	1.078	0.240	<0.001	2.94	<0.001
OCUPADO	1.027	0.198	<0.001	2.79	<0.001
ACTITUD	1.105	0.189	<0.001	3.02	<0.001
VISION	1.612	0.300	<0.001	5.01	<0.001
RIESGO	0.774	0.201	<0.001	2.17	<0.001
CREATIV	0.589	0.211	0.005	1.80	0.005
CLUSTER**					<0.001

* Probabilidad asociada a los estadísticos del cociente de verosimilitud para pruebas de tipo III.
** La variable CLUSTER tiene 13 grados de libertad.

Como puede observarse en la tabla anterior, la totalidad de las covariables, incluyendo CLUSTER, son estadísticamente significativas. No obstante, este modelo es excluido de los análisis posteriores por las razones antes esbozadas.

Interacciones dobles

Si bien resulta deseable mantener al modelo tan parsimonioso como sea posible, es necesario evaluar si la incorporación de términos de interacción doble en el predictor lineal mejora el ajuste. En este caso, únicamente dos interacciones se han revelado como estadísticamente significativas[31]:

- GENERO*CREATIV
- OCUPADO*RIESGO

[31] Un efecto de interacción doble significativo implica que el efecto de una covariable sobre la variable respuesta –o una función de la misma–, difiere de acuerdo al valor que asuma la segunda covariable.

Es interesante comentar el significado de las mismas. Para ello, se ejemplifica con los coeficientes estimados bajo el modelo mixto con verosimilitud completa (NLMIXED) por contarse en tales casos con medidas de bondad del ajuste sencillas de implementar y comparar entre sí. Para el término GENERO*CREATIV, los cocientes de chances estimados exhibidos en la Tabla 32 abarcan las cuatro combinaciones posibles: *(a)* hombre con alta creatividad; *(b)* hombre con baja creatividad; *(c)* mujer con alta creatividad; *(d)* mujer con baja creatividad.

TABLA 32: Coeficientes estimados para la interacción GENERO*CREATIV

Interacción	exp(β)
GENERO = 1; CREATIV = 1	**4.5**
GENERO = 1; CREATIV = 0	3.6
GENERO = 0; CREATIV = 1	3.1
GENERO = 0; CREATIV = 0	1.0

Tal como surge de la tabla anterior, considerando que el efecto aleatorio es igual a cero y controlando por las restantes covariables, respecto de una mujer con baja creatividad, las chances de que un alumno posea vocación emprendedora se multiplican: *(i)* por un factor de 4.5 si éste es un hombre con alta creatividad; *(ii)* por un factor de 3.6 si se trata de un hombre con baja creatividad; *(iii)* por un factor de 3.1 si se trata de una mujer altamente creativa. Es decir, tanto la condición de ser hombre como de tener creatividad elevada incrementan las chances de que el alumno posea vocación emprendedora (Figura 3).

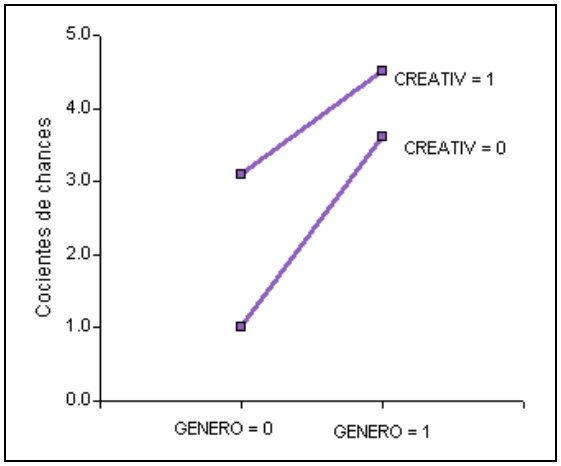

FIGURA 3: Efecto estimado de la interacción GENERO*CREATIV sobre la probabilidad de vocación emprendedora

Con respecto a la interacción OCUPADO*RIESGO, los cocientes de chances estimados incluidos en la Tabla 33 se refieren a las combinaciones: *(a)* ocupado propenso al riesgo; *(b)* ocupado adverso al riesgo; *(c)* desocupado/inactivo propenso al riesgo; *(d)* desocupado/inactivo adverso al riesgo.

TABLA 33: Coeficientes estimados para la interacción OCUPADO*RIESGO

Interacción	exp(β)
OCUPADO = 1; RIESGO = 1	**6.8**
OCUPADO = 1; RIESGO = 0	2.1
OCUPADO = 0; RIESGO = 1	1.3
OCUPADO = 0; RIESGO = 0	1.0

Controlando por las restantes covariables y suponiendo nulo el efecto aleatorio, en relación a un individuo que no está ocupado y es adverso al riesgo, las chances de que un alumno posea vocación emprendedora se multiplican: *(i)* por un factor de 6.8 si éste se encuentra ocupado y es propenso al riesgo; *(ii)* se duplican si está ocupado pero es adverso al riesgo; *(iii)* se incrementan sólo un 30% si el alumno es propenso al riesgo pero no se halla trabajando. En consecuencia, la combinación de que un individuo trabaje y sea tomador de riesgo incrementa notablemente las chances de poseer vocación emprendedora, más allá del efecto que ejercen por separado cada una de estas condiciones (Figura 4).

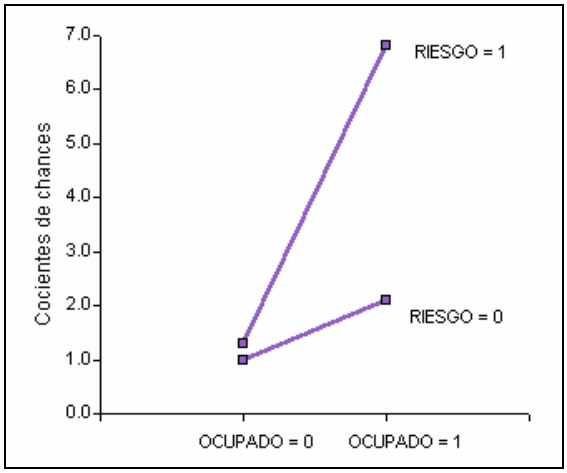

FIGURA 4: Efecto estimado de la interacción OCUPADO*RIESGO sobre la probabilidad de vocación emprendedora

Estrictamente hablando, las interacciones representan la falta de aditividad en la escala *logit*. Para evaluar si es conveniente incorporar estos términos al predictor lineal, se comparan los valores de AIC y de BIC (Tabla 34), observándose que los valores de las medidas de bondad del ajuste disminuyen al adicionar en el modelo las interacciones dobles.

TABLA 34: Estadísticos de bondad del ajuste correspondientes al modelo mixto con verosimilitud completa con y sin interacciones dobles

	AIC	BIC
MODELO DE EFECTOS PRINCIPALES	**757.1**	**762.2**
GENERO*CREATIV	754.6	760.3
OCUPADO*RIESGO	754.3	760.0
AMBAS INTERACCIONES	751.2	757.5

Dado que las dos interacciones dobles analizadas son ordenadas –i.e., los efectos no se cruzan–, lo que cambia es la magnitud pero no el sentido de la asociación entre las variables. Por lo tanto, como no se modifica la interpretación de los efectos principales, se opta por un modelo más parsimonioso que no incluya los términos de interacción.

7.3. Inferencia

Antes de avanzar en el análisis, es importante comparar los estimadores de los parámetros y de los errores estándares obtenidos por los distintos ajustes. Con ello se va a intentar responder si las estimaciones de los efectos dependen de la forma de modelar la dependencia entre las observaciones y si hay algún enfoque que brinde las estimaciones más eficientes.

Primero, en las Tablas 35 y 36 se comparan los valores de los parámetros estimados, calculándose los cocientes entre los coeficientes obtenidos mediante las distintas modalidades de análisis. Se excluye de este análisis a la estructura de logaritmos de cocientes de chances por bloque (LOGOR=LOGORVAR), por no diferir en forma sustancial de los otros ajustes. En la primera fila del encabezado de cada tabla aparece la estrategia que corresponde al numerador y, en la segunda fila, la que corresponde al denominador del cociente. Los cocientes resaltados son aquéllos que evidencian una diferencia mayor o igual al 10% entre las estimaciones de los distintos ajustes.

TABLA 35: Cocientes entre parámetros estimados bajo el enfoque marginal

Covariable	Independencia			Corr. intercambiable		Log OR intercam-biables
	Corr. intercam-biable	Log OR intercam-biables	Log OR anidados a 1 nivel	Log OR intercam-biables	Log OR anidados a 1 nivel	Log OR anidados a 1 nivel
INTERCEPTO	1.03	1.01	1.01	0.98	0.98	1.00
GENERO	**0.86**	**0.88**	**0.88**	1.03	1.02	0.99
OCUPADO	1.05	1.03	1.03	0.98	0.98	1.00
ACTITUD	1.03	1.02	1.02	0.99	0.99	1.00
VISION	1.07	1.03	1.03	0.96	0.96	1.00
RIESGO	1.07	1.05	1.05	0.99	0.99	1.00
CREATIV	0.93	0.95	0.93	1.02	1.00	0.98

TABLA 36: Cocientes entre parámetros estimados bajo los enfoques marginal y mixto

Covariable	Independencia		Corr. intercambiable		Log OR intercambiables	
	Verosim. completa	Verosim. condicional	Verosim. completa	Verosim. condicional	Verosim. completa	Verosim. condicional
INTERCEPTO	1.00		0.97		0.98	
GENERO	**0.81**	**0.76**	0.94	**0.89**	0.92	**0.87**
OCUPADO	1.02	1.05	0.97	1.00	0.99	1.02
ACTITUD	0.99	1.00	0.96	0.96	0.97	0.97
VISION	1.00	1.01	0.94	0.95	0.98	0.98
RIESGO	1.04	1.08	0.97	1.01	0.99	1.03
CREATIV	**0.89**	**0.85**	0.95	0.91	0.93	**0.89**

TABLA 36 (CONTINUACIÓN): Cocientes entre parámetros estimados bajo los enfoques marginal y mixto

Covariable	Log OR anidados a 1 nivel		Verosim. completa
	Verosim. completa	Verosim. condicional	Verosim. condicional
INTERCEPTO	0.99		
GENERO	0.93	**0.87**	0.94
OCUPADO	0.99	1.02	1.03
ACTITUD	0.97	0.97	1.01
VISION	0.98	0.98	1.01
RIESGO	0.99	1.03	1.04
CREATIV	0.95	0.90	0.95

Tal como se desprende de las tablas previas, los parámetros estimados exhiben mínimas variaciones para las distintas estructuras de dependencia (Tabla 35), así como entre los enfoques marginal y mixto (Tabla 36). No obstante, se observa que:

- El efecto asociado a la variable GENERO sobre la probabilidad de que el alumno posea vocación emprendedora, es menor si se opta por la estructura de independencia (TYPE=IND).
- El efecto asociado a las variables GENERO y CREATIV sobre la probabilidad de que el alumno posea vocación emprendedora, es menor si se emplea el enfoque marginal

que el mixto. I.e., es menor para el promedio poblacional que para un *cluster* específico.

- Para la estructura de correlación y de logaritmos de cocientes de chances intercambiable (TYPE=EXCH y LOGOR=EXCH), los parámetros estimados con el enfoque marginal son levemente menores, en valor absoluto, a los obtenidos con el modelo mixto con verosimilitud completa.

A fin de explorar la precisión de las estimaciones, se analiza la amplitud de los intervalos de confianza correspondientes a los coeficientes estimados bajo los distintos ajustes. Para poder comparar los valores entre sí, la amplitud del intervalo se divide por el estimador puntual, calculándose así una amplitud relativa (Tabla 37).

TABLA 37: Amplitud relativa de los intervalos de confianza para β

Covariable	Obs. indepen-dientes	Modelos marginales				Modelos mixtos	
		Indepen-dencia	Corr. intercam-biable	Log OR intercam-biables	Log OR anidados a 1 nivel	Verosim. completa	Verosim. condicional
GENERO	0.99	0.89	0.64	0.68	0.68	1.01	0.89
OCUPADO	0.69	0.90	0.92	0.90	0.89	0.81	0.77
ACTITUD	0.66	0.47	0.49	0.48	0.48	0.57	0.68
VISION	0.70	0.80	0.75	0.80	0.79	0.79	0.73
RIESGO	0.92	0.83	0.86	0.86	0.86	1.08	1.03
CREATIV	1.63	1.87	1.66	1.72	1.70	1.64	1.50
Media	**0.93**	**0.96**	**0.89**	**0.91**	**0.90**	**0.98**	**0.93**
Coeficiente de variación	**39.65%**	**49.52%**	**46.00%**	**47.22%**	**46.39%**	**37.55%**	**32.61%**

En las dos últimas filas de la tabla anterior se reportan la media y el coeficiente de variación de la amplitud relativa para cada ajuste. En base a dichos estadísticos se concluye que:

- Los intervalos más estrechos son, en promedio, los que se obtienen bajo el modelo marginal con estructura de correlación de simetría compuesta (TYPE=EXCH).
- Los intervalos más amplios son, en promedio, los que se obtienen bajo el modelo mixto con verosimilitud completa (NLMIXED).
- La amplitud relativa es menos variable bajo los modelos de regresión logística ordinaria y mixtos, y más variable bajo los modelos marginales.

Otro aspecto interesante de analizar es la eficiencia de los estimadores en términos de los errores estándares estimados. Para ello, se ha calculado como medida de eficiencia relativa el cociente entre los errores estándares obtenidos bajo los distintos ajustes (Tablas 38 y 39). Nuevamente, en la primera fila del encabezado de cada tabla aparece la estrategia cuyo error estándar corresponde al numerador y, en la segunda fila, la que corresponde al denominador del cociente. Los valores resaltados son aquéllos que evidencian una diferencia mayor o igual al 20% entre las respectivas estimaciones.

TABLA 38: Eficiencia relativa bajo el enfoque marginal

Covariable	Independencia			Corr. intercambiable		Log OR intercam-biables
	Corr. intercam-biable	Log OR intercam-biables	Log OR anidados a 1 nivel	Log OR intercam-biables	Log OR anidados a 1 nivel	Log OR anidados a 1 nivel
INTERCEPTO	**1.22**	1.16	1.15	0.96	0.95	0.99
GENERO	1.19	1.15	1.13	0.97	0.95	0.98
OCUPADO	1.04	1.04	1.05	1.01	1.01	1.01
ACTITUD	0.98	0.99	1.00	1.01	1.02	1.01
VISION	1.14	1.04	1.04	0.91	0.91	1.00
RIESGO	1.03	1.01	1.01	0.99	0.99	1.00
CREATIV	1.05	1.03	1.03	0.98	0.98	1.00

Los errores estándares estimados bajo el enfoque marginal, para las diversas alternativas elegidas para modelar la dependencia entre las observaciones, prácticamente no difieren entre sí. En otras palabras, la estructura seleccionada para la matriz de correlación de trabajo no altera sustancialmente la eficiencia. Este resultado es satisfactorio, ya que permite corroborar que la estimación es robusta, lo cual no se verificaría si las diferencias fuesen notorias. No obstante, los mayores valores son los que se obtienen con la estructura de independencia (TYPE=IND).

Al comparar los modelos marginales con los modelos mixtos, se evidencia una mayor variabilidad entre los errores estándares estimados que la observada entre los coeficientes. En todos los casos, con el enfoque marginal se obtiene una estimación más eficiente para las variables GENERO y ACTITUD, y en menor medida para RIESGO, mientras que los enfoques mixtos estiman más eficientemente al parámetro de la variable OCUPADO. A su vez, las estimaciones son igualmente eficientes si se obtienen mediante el modelo mixto con verosimilitud completa o con verosimilitud condicional.

TABLA 39: Eficiencia relativa bajo los enfoques marginal y mixto

Covariable	Independencia		Corr. intercambiable		Log OR intercambiables	
	Verosim. completa	Verosim. condicional	Verosim. completa	Verosim. condicional	Verosim. completa	Verosim. condicional
INTERCEPTO	1.08		0.89		0.93	
GENERO	**0.79**	**0.76**	**0.66**	**0.64**	**0.68**	**0.66**
OCUPADO	**1.26**	**1.24**	**1.21**	1.20	**1.21**	1.19
ACTITUD	**0.69**	**0.68**	**0.70**	**0.70**	**0.69**	**0.69**
VISION	1.13	1.11	0.99	0.97	1.09	1.07
RIESGO	0.88	0.87	0.85	0.85	0.86	0.86
CREATIV	1.11	1.11	1.06	1.05	1.08	1.07

TABLA 39 (CONTINUACIÓN): Eficiencia relativa bajo los enfoques marginal y mixto

Covariable	Log OR anidados a 1 nivel		Verosim. completa
	Verosim. completa	Verosim. condicional	Verosim. condicional
INTERCEPTO	0.94		
GENERO	**0.69**	**0.67**	0.97
OCUPADO	**1.20**	1.18	0.99
ACTITUD	**0.69**	**0.68**	1.00
VISION	1.09	1.07	0.98
RIESGO	0.86	0.86	0.99
CREATIV	1.08	1.07	1.00

Por último, en la Tabla 40 se comparan los cocientes entre los errores estándares que surgen de una regresión logística ordinaria (sin efectos fijos de *cluster*) con aquéllos obtenidos bajo los restantes ajustes.

TABLA 40: Eficiencia relativa con y sin independencia en las observaciones

Covariable	Observaciones independientes					
	Indepen-dencia	Corr. simetría compuesta	Log OR simetría compuesta	Log OR anidados a 1 nivel	Verosim. completa	Verosim. condicional
INTERCEPTO	**0.80**	0.97	0.93	0.92	0.86	
GENERO	1.12	**1.33**	**1.29**	**1.27**	0.88	0.85
OCUPADO	**0.77**	**0.79**	**0.80**	0.80	0.96	0.95
ACTITUD	**1.41**	**1.39**	**1.40**	**1.41**	0.97	0.97
VISION	0.87	0.99	0.90	0.90	0.98	0.96
RIESGO	1.11	1.14	1.12	1.12	0.97	0.96
CREATIV	0.87	0.92	0.90	0.90	0.97	0.97

Relacionando al modelo de regresión logística ordinaria con el enfoque marginal, se aprecia que:

- El efecto asociado a la variable OCUPADO sobre la probabilidad de que el alumno posea vocación emprendedora, dadas las restantes covariables, es más

eficientemente estimado si se ignora la dependencia existente al interior de los *clusters.*

- Los efectos de las variables GENERO y ACTITUD, controlando por las demás covariables, se estiman con mayor eficiencia si se incorpora en el modelo la dependencia entre las respuestas.

Al comparar el modelo con observaciones independientes con los modelos mixtos, se observa que no hay diferencias relevantes entre los errores estándares estimados. No obstante, los hallados mediante una regresión logística ordinaria son apenas menores que los estimados bajo los modelos con verosimilitud completa y condicional.

En la comparación de la eficiencia brindada por los distintos ajustes no se han encontrado diferencias relevantes, lo que se explica por la baja correlación entre las respuestas. Sin embargo, debe tenerse presente que la obtención de errores estándares más pequeños no es el único objetivo que se persigue al contemplar la dependencia entre las observaciones.

En la elección del modelo a estimar está implícito el tipo de inferencia que se quiere efectuar, lo cual justifica que se utilice un determinado enfoque más allá de la eficiencia. Si desean hallarse coeficientes que representen un promedio para todos los *clusters* y que la inferencia corresponda al promedio poblacional *(population average inference)*, debe optarse por el enfoque marginal. En este caso, el método de ecuaciones de estimación generalizadas garantiza que los errores estándares se estimen en forma robusta.

Por su parte, si desea "controlarse" por el efecto de *cluster* y que los parámetros estimados posean una interpretación específica para cada *cluster (cluster specific inference)*, el enfoque a elegir es el mixto. En este caso, el costo será que los errores estándares sean quizás más grandes, dado que el espacio de inferencia es más amplio. Si hay mayor variabilidad entre *clusters* que dentro de los mismos –lo que en última instancia justifica que el muestreo sea por conglomerados–, se obtendrá una mayor variabilidad entre los errores estándares estimados con el modelo mixto que con el modelo marginal.

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, cuya configuración se exhibe en la Tabla 41, surgen de la tabla de contingencia parcial a cinco vías de clasificación que es observada con mayor frecuencia[32]:

> La 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.

TABLA 41: Configuración de la categoría modal

Covariable	Niveles	Categoría modal
GENERO	Hombre	X
	Mujer	
OCUPADO	Ocupado	X
	Desocupado / Inactivo	
ACTITUD	Actitud empresarial frente al desempleo	
	Actitud no empresarial frente al desempleo	X
VISION	Visión favorable de la actividad empresarial	X
	Visión desfavorable de la actividad empresarial	
RIESGO	Propenso al riesgo	
	Adverso al riesgo	X
CREATIV	Alta creatividad	
	Baja creatividad	X

Este apartado está organizado de la siguiente manera. Primero se estima la probabilidad de VE=1 para un individuo perteneciente a la categoría modal bajo el enfoque marginal y se calculan coeficientes de riesgo relativo entre los dos niveles de cada una de las covariables (Tabla 42). Luego, se calcula la probabilidad de VE=1 y los riesgos relativos bajo el modelo mixto con verosimilitud completa, controlándose por el efecto que ejerce la facultad a la que asiste el alumno (Tablas 43 y 44). Seguidamente, se repiten los cálculos para el modelo mixto

con verosimilitud condicional, utilizando como aproximación al intercepto los valores hallados aplicando la expresión [*76*] (Tablas 45 y 46). Por último, se obtienen las estimaciones para una regresión logística ordinaria que ignora la dependencia entre las respuestas.

A partir de las estimaciones efectuadas, la probabilidad de que un alumno posea vocación emprendedora bajo el enfoque marginal, modelando la dependencia mediante la estructura de correlación de simetría compuesta (TYPE=EXCH), resulta:

$$\hat{\mu}_{ij} = \frac{\exp\left(-3.80 + 0.93GENERO_{ij} + 1.00OCUPADO_{ij} + 1.04ACTITUD_{ij} + 1.50VISION_{ij} + 0.77RIESGO_{ij} + 0.52CREATIV_{ij}\right)}{1 + \exp\left(-3.80 + 0.93GENERO_{ij} + 1.00OCUPADO_{ij} + 1.04ACTITUD_{ij} + 1.50VISION_{ij} + 0.77RIESGO_{ij} + 0.52CREATIV_{ij}\right)}.$$

La probabilidad de que un alumno perteneciente a la categoría modal posea vocación emprendedora se estima en **0.41** (Tabla 42). 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 42: Probabilidades estimadas y riesgos relativos bajo el modelo marginal con estructura de correlación de simetría compuesta (TYPE=EXCH)

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

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 a las otras covariables. Lo que más impacta sobre la vocación emprendedora son:

[32] Cada tabla combina como vías de clasificación: GENERO, OCUPADO, ACTITUD, VISION y RIESGO, cruzándose al interior CREATIV con VE. Al ser todas las variables dicotómicas ello da lugar a $32 = 2^5$ combinaciones posibles. En otras

- La actitud frente al desempleo: la probabilidad de que un alumno posea vocación emprendedora es un 62% mayor si, ante la falta de un trabajo adecuado a su formación profesional al momento de graduarse, asume una actitud empresarial.
- La visión de la actividad empresarial: un alumno que tiene una visión desfavorable de la actividad emprendedora tiene un 67% menos de probabilidad de poseer vocación emprendedora.

Bajo el enfoque marginal, la probabilidad de que un individuo con todas las características positivamente asociadas con la vocación emprendedora efectivamente tenga vocación emprendedora se estima en **0.88**, correspondiendo este valor a un promedio poblacional. En otras palabras:

> Se estima que un alumno hombre, ocupado, con una actitud emprendedora frente al desempleo, que tenga una visión favorable de la actividad empresarial, sea propenso al riego y altamente creativo, tiene una probabilidad de 0.88 de poseer vocación emprendedora.

Al respecto, debe enfatizarse que es muy difícil explicar con pocas covariables la vocación emprendedora de un individuo, al ser esta condición el resultado de un conjunto de elementos que ejercen influencia a lo largo de su vida, de características personales innatas y de factores situacionales.

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 2. En este modelo, la probabilidad condicional se calcula como:

$$\hat{\mu}_{ij}/U_i = \frac{\exp\left(-3.94 + 0.99GENERO_{ij} + 1.04OCUPADO_{ij} + 1.09ACTITUD_{ij} + 1.60VISION_{ij} + 0.79RIESGO_{ij} + 0.55CREATIV_{ij} + U_i\right)}{1 + \exp\left(-3.94 + 0.99GENERO_{ij} + 1.04OCUPADO_{ij} + 1.09ACTITUD_{ij} + 1.60VISION_{ij} + 0.79RIESGO_{ij} + 0.55CREATIV_{ij} + U_i\right)}.$$

En la última columna de la Tabla 43 se indican los cocientes entre las probabilidades estimadas para un alumno perteneciente a la categoría modal de cada una de las facultades y

palabras, el cuerpo de datos puede pensarse como una tabla de contingencia 2x2x2x2x2.

la probabilidad asociada a la categoría modal considerando que el efecto aleatorio es nulo. La probabilidad de que un alumno perteneciente a la categoría modal posea vocación emprendedora es de **0.42** en una facultad típica.

TABLA 43: Efecto de la facultad sobre la probabilidad de presencia de vocación emprendedora bajo el modelo mixto con verosimilitud completa (NLMIXED)

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

En comparación con una facultad típica, es al menos un 20% más probable que posea vocación emprendedora si concurre a las facultades[33]:

- U3 → privada, zona 1, economía y administración.
- U4 → privada, zona 1, economía y administración.
- U12 → privada, zona 2, ingeniería.
- U7 → privada, zona 1, economía y administración.

El efecto asociado a la U3 es particularmente alto, observándose que un alumno con las características de la categoría modal tiene un 50% más de probabilidad de poseer vocación emprendedora si estudia en esa institución. Por el contrario, la probabilidad estimada es al menos un 20% menor de 0.42 si el alumno asiste a las facultades:

- U9 → pública, zona 2, ingeniería.
- U5 → pública, zona 2, economía y administración.
- U14 → pública, zona 2, ingeniería.

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 en la

sección 6 aplicando inferencia clásica. En cuanto al efecto de la localización geográfica, estudios en la temática de creación de empresas reportan una alta correlación positiva entre el tamaño de la localidad y la tasa de creación de empresas que en ella se evidencia (Gennero *et al.*, 1999). Una ciudad más grande, entre otras cosas, ofrece un mercado más atractivo para el emprendedor.

En la Tabla 44 se calcula el riesgo relativo para cada facultad, i.e., contemplando el efecto aleatorio en el predictor lineal. Para ello se dividen la probabilidad condicional de VE=1 si las características del alumno no coinciden con la categoría modal y la probabilidad condicional de VE=1 si el alumno pertenece a dicha categoría.

TABLA 44: Riesgos relativos por facultad
bajo el modelo mixto con verosimilitud completa (NLMIXED)

Nivel covariables	Riesgo relativo				
	U1	U2	U3	U4	U5
✓ Si el alumno es mujer	0.54	0.48**	0.62*	0.58	0.46**
✓ Si el alumno no está trabajando	0.52	0.46	0.60*	0.56*	0.44
✓ Si el alumno posee actitud empresarial frente al desempleo	1.50	1.76	1.33**	1.39**	1.86*
✓ Si el alumno posee una visión desfavorable de la actividad empresarial	0.33	0.28	0.40*	0.37*	0.27*
✓ Si el alumno es propenso al riesgo	1.38	1.55	1.26**	1.30**	1.62
✓ Si el alumno posee alta creatividad	1.27	1.38	1.19	1.22	1.42*

* Indica que el cociente en la facultad analizada es un 10% mayor que el promedio poblacional.
** Indica que el cociente en la facultad analizada es un 10% menor que el promedio poblacional.

TABLA 44 (CONTINUACIÓN): Riesgos relativos por facultad
bajo el modelo mixto con verosimilitud completa (NLMIXED)

Nivel covariables	Riesgo relativo				
	U6	U7	U8	U9	U10
✓ Si el alumno es mujer	0.49	0.55	0.52	0.44**	0.49
✓ Si el alumno no está trabajando	0.47	0.53	0.50	0.42**	0.47
✓ Si el alumno posee actitud empresarial frente al desempleo	1.70	1.48	1.57	1.98*	1.70
✓ Si el alumno posee una visión desfavorable de la actividad empresarial	0.29	0.34	0.32	0.25**	0.29
✓ Si el alumno es propenso al riesgo	1.51	1.36	1.43	1.69*	1.51
✓ Si el alumno posee alta creatividad	1.36	1.26	1.30	1.46*	1.35

* Indica que el cociente en la facultad analizada es un 10% mayor que el promedio poblacional.
** Indica que el cociente en la facultad analizada es un 10% menor que el promedio poblacional.

[33] Zona 1 = Ciudad Autónoma de Buenos Aires y Gran Buenos Aires; Zona 2 = resto de la Provincia de Buenos Aires.

TABLA 44 (CONTINUACIÓN): Riesgos relativos por facultad bajo el modelo mixto con verosimilitud completa (NLMIXED)

Nivel covariables	Riesgo relativo			
	U11	U12	U13	U14
✓ Si el alumno es mujer	0.51	0.55	0.49	0.47**
✓ Si el alumno no está trabajando	0.49	0.53	0.47	0.45
✓ Si el alumno posee actitud empresarial frente al desempleo	1.63	1.47	1.72	1.80*
✓ Si el alumno posee una visión desfavorable de la actividad empresarial	0.30	0.34	0.29	0.27**
✓ Si el alumno es propenso al riesgo	1.46	1.36	1.52	1.58
✓ Si el alumno posee alta creatividad	1.32	1.26	1.36	1.39

* Indica que el cociente en la facultad analizada es un 10% mayor que el promedio poblacional.
** Indica que el cociente en la facultad analizada es un 10% menor que el promedio poblacional.

En general, estos cocientes no difieren de los hallados bajo el enfoque marginal. Sin embargo, en las U3, U4, U5, U9 y U14, algunos valores estimados distan en más de un 10%, en valor absoluto, de los obtenidos para el promedio de la población. La interpretación de estos resultados, ejemplificando con la U3, implica que:

> Para un alumno de la U3 que difiere de la categoría modal por ser mujer, por no estar trabajando actualmente o por carecer de una actitud emprendedora frente al desempleo, el impacto sobre la probabilidad de poseer vocación emprendedora es menor que para el promedio poblacional. Sin embargo, este impacto es mayor que para el promedio de la población si valora negativamente la actividad empresarial o es adverso al riego.

Reiterando este análisis para cada una de las covariables –i.e., comparando el efecto del cambio en el nivel de una covariable desde la categoría modal en una cierta facultad en relación con el mismo cambio para el promedio poblacional–, resulta que:

- Ser mujer impacta menos sobre la probabilidad de VE=1 en la U3 e impacta más en las U5, U9 y U14.

- No estar ocupado impacta menos sobre la probabilidad de VE=1 en las U3 y U4, mientras que impacta más en la U9.
- Carecer de actitud empresarial frente al desempleo impacta menos sobre la probabilidad de VE=1 en las U5, U9 y U14, mientras que impacta más en las U3 y U4.
- Poseer una visión desfavorable de la actividad empresarial impacta menos sobre la probabilidad de VE=1 en las U3, U4 y U5, e impacta más en las U9 y U14.
- Ser propenso al riesgo impacta menos sobre la probabilidad de VE=1 en la U9 e impacta más en las U3 y U4.
- Poseer un alto nivel de creatividad impacta menos sobre la probabilidad de VE=1 en las U5 y U9.

Con respecto al enfoque condicional, éste no permite estimar las medias individuales al condicionar al intercepto fuera del modelo. Sin embargo, si se aproxima este valor desconocido por los valores obtenidos en la Tabla 28, es posible estimar la probabilidad condicional de VE=1 para un alumno de una facultad determinada. Cabe aclarar que sigue siendo imposible hacerlo para una facultad típica en la cual el efecto aleatorio sea nulo. Al predictor lineal estimado:

$$\hat{\eta}_{ij} = 1.05 GENERO_{ij} + 1.03 OCUPADO_{ij} + 1.08 ACTITUD_{ij} + 1.59 VISION_{ij} + 0.76 RIESGO_{ij} + 0.57 CREATIV_{ij} .$$

se le debe sumar una constante que asume distintos valores en cada facultad, como paso previo al cálculo de las probabilidades (Tabla 45).

Si estos valores se comparan con los obtenidos bajo el modelo mixto con verosimilitud completa (Tabla 43), se observa que, exceptuando a las U3 y U4, las probabilidades estimadas son inferiores. Aunque los parámetros estimados no difieren significativamente bajo ambos ajustes, la diferencia se explica porque $\exp\left(\hat{\beta}_0\right)$ es, para los restantes *clusters*, mayor si se estima con el modelo mixto con verosimilitud completa.

TABLA 45: Probabilidad estimada bajo el modelo mixto con verosimilitud condicional (PHREG)

Categoría modal	$Pr(VE=1/U_i)$
✓ U1	0.39
✓ U2	0.23
✓ U3	0.81
✓ U4	0.60
✓ U5	0.21
✓ U6	0.29
✓ U7	0.47
✓ U8	0.46
✓ U9	0.13
✓ U10	0.27
✓ U11	0.27
✓ U12	0.45
✓ U13	0.27
✓ U14	0.22

Repitiendo el análisis efectuado para el modelo mixto con verosimilitud completa (Tabla 44), se incluyen en la Tabla 46 los riesgos relativos calculados como el cociente entre las probabilidades de VE=1 para un alumno en la modalidad de interés y en la categoría modal. Dado que no se evidencian mayores diferencias entre ambos enfoques mixtos, el modelo con verosimilitud condicional se excluye de los análisis que se efectúan en adelante.

TABLA 46: Riesgos relativos por facultad bajo el modelo mixto con verosimilitud condicional (PHREG)

Nivel covariables	Riesgo relativo				
	U1	U2	U3	U4	U5
✓ Si el alumno es mujer	0.47**	0.41**	0.74*	0.57	0.41**
✓ Si el alumno no está trabajando	0.49	0.43	0.75*	0.59*	0.43**
✓ Si el alumno posee actitud empresarial frente al desempleo	1.67	2.04*	1.15**	1.37**	2.07*
✓ Si el alumno posee una visión desfavorable de la actividad empresarial	0.30	0.25**	0.57*	0.39*	0.25*
✓ Si el alumno es propenso al riesgo	1.48	1.70*	1.11**	1.28**	1.71*
✓ Si el alumno posee alta creatividad	1.36	1.51*	1.09**	1.22	1.52*

* Indica que el cociente en la facultad analizada es un 10% mayor que el promedio poblacional.
** Indica que el cociente en la facultad analizada es un 10% menor que el promedio poblacional.

TABLA 46 (CONTINUACIÓN): Riesgos relativos por facultad bajo el modelo mixto con verosimilitud condicional (PHREG)

Nivel covariables	Riesgo relativo				
	U6	U7	U8	U9	U10
✓ Si el alumno es mujer	0.43**	0.51	0.50	0.38**	0.43**
✓ Si el alumno no está trabajando	0.45	0.52	0.52	0.40**	0.44
✓ Si el alumno posee actitud empresarial frente al desempleo	1.87*	1.53	1.56	2.36*	1.92*
✓ Si el alumno posee una visión desfavorable de la actividad empresarial	0.27**	0.33	0.32	0.23**	0.26**
✓ Si el alumno es propenso al riesgo	1.60	1.39	1.40	1.86*	1.63*
✓ Si el alumno posee alta creatividad	1.44*	1.30	1.31	1.61*	1.47*

* Indica que el cociente en la facultad analizada es un 10% mayor que el promedio poblacional.
** Indica que el cociente en la facultad analizada es un 10% menor que el promedio poblacional.

TABLA 46 (CONTINUACIÓN): Riesgos relativos por facultad bajo el modelo mixto con verosimilitud condicional (PHREG)

Nivel covariables	Riesgo relativo			
	U11	U12	U13	U14
✓ Si el alumno es mujer	0.42**	0.50	0.43**	0.41**
✓ Si el alumno no está trabajando	0.44	0.51	0.45	0.43**
✓ Si el alumno posee actitud empresarial frente al desempleo	1.94*	1.57	1.90*	2.07*
✓ Si el alumno posee una visión desfavorable de la actividad empresarial	0.26**	0.32	0.26**	0.25**
✓ Si el alumno es propenso al riesgo	1.64*	1.41	1.62	1.71*
✓ Si el alumno posee alta creatividad	1.47*	1.31	1.46*	1.52*

* Indica que el cociente en la facultad analizada es un 10% mayor que el promedio poblacional.
** Indica que el cociente en la facultad analizada es un 10% menor que el promedio poblacional.

Finalmente, se estima la probabilidad de que un alumno perteneciente a la categoría modal posea vocación emprendedora bajo un modelo de regresión logística ordinaria. Debe tenerse en claro que en este apartado se han realizado estimaciones puntuales y que, precisamente, la diferencia entre este modelo y un modelo marginal con estructura de independencia (TYPE=IND) reside en los errores estándares estimados. Por lo tanto, la estimación puntual corresponde tanto a uno como a otro tipo de ajuste.

La probabilidad estimada se deriva de la siguiente expresión:

$$\hat{\mu}_{ij} = \frac{\exp\left(-3.92 + 0.80GENERO_{ij} + 1.06OCUPADO_{ij} + 1.07ACTITUD_{ij} + 1.60VISION_{ij} + 0.82RIESGO_{ij} + 0.48CREATIV_{ij}\right)}{1 + \exp\left(-3.92 + 0.80GENERO_{ij} + 1.06OCUPADO_{ij} + 1.07ACTITUD_{ij} + 1.60VISION_{ij} + 0.82RIESGO_{ij} + 0.48CREATIV_{ij}\right)},$$

y, para un individuo modal, se estima en **0.39**, valor algo inferior al obtenido por los otros dos enfoques. Aún resta por ver la precisión con la que tales probabilidades han sido estimadas, lo que puede apreciarse en la Tabla 47.

En este análisis se contemplan dos opciones para tratar al efecto aleatorio del modelo de verosimilitud completa: *(a)* igualarlo a cero, realizando inferencia para un *cluster* típico; *(b)* suponerlo equivalente al promedio ponderado de los efectos aleatorios estimados por *cluster* (–0.1277). En este último caso, el estimador puntual de la proporción de alumnos pertenecientes a la categoría modal que poseen vocación emprendedora es inferior.

TABLA 47: Estimación puntual y amplitud del intervalo de confianza de la probabilidad de vocación emprendedora bajo los distintos ajustes

Categoría modal	Estimación puntual	Amplitud intervalo de confianza
Modelo marginal con estructura de correlación intercambiable (TYPE=EXCH)	0.409	0.176
Modelo mixto con verosimilitud completa (NLMIXED) suponiendo el efecto aleatorio nulo	0.421	0.259
Modelo mixto con verosimilitud completa (NLMIXED) suponiendo el efecto aleatorio promedio	0.390	0.253
Regresión logística ordinaria	0.387	0.158

Comparando los ajustes, se observa que el intervalo de confianza es apenas más estrecho si se considera que las respuestas son independientes que si se ajusta el modelo marginal. Sin embargo, ya se ha explicado que éste no es el único objetivo que se persigue al ajustar un modelo que incorpore la dependencia entre las observaciones. Son aspectos muy importantes que los errores estándares de los parámetros se estimen en forma robusta y que el modelo ofrezca alta predictibilidad. Y, fundamentalmente, la elección del enfoque depende del tipo de inferencia que desea realizarse, así como de las preguntas de investigación que se quieren responder.

7.4. Diagnóstico

Para evaluar la calidad del ajuste de los modelos formulados, existen pocas herramientas disponibles debido a dos motivos:

- La naturaleza binaria de la variable respuesta y de las covariables restringe las técnicas formales y gráficas a aplicar.
- Aún no se han desarrollado técnicas formales adecuadas a los modelos marginales basados en la función de cuasi-verosimilitud.

Si bien los residuos de Anscombe se destacaron por sus propiedades distribucionales como los más apropiados para diagnosticar un modelo con observaciones binomiales, éstos no resultan adecuados si la variable respuesta posee distribución binaria. A esta conclusión se llega luego de analizar los correspondientes histogramas y *Q-Q plots* (Figuras 5 y 6).

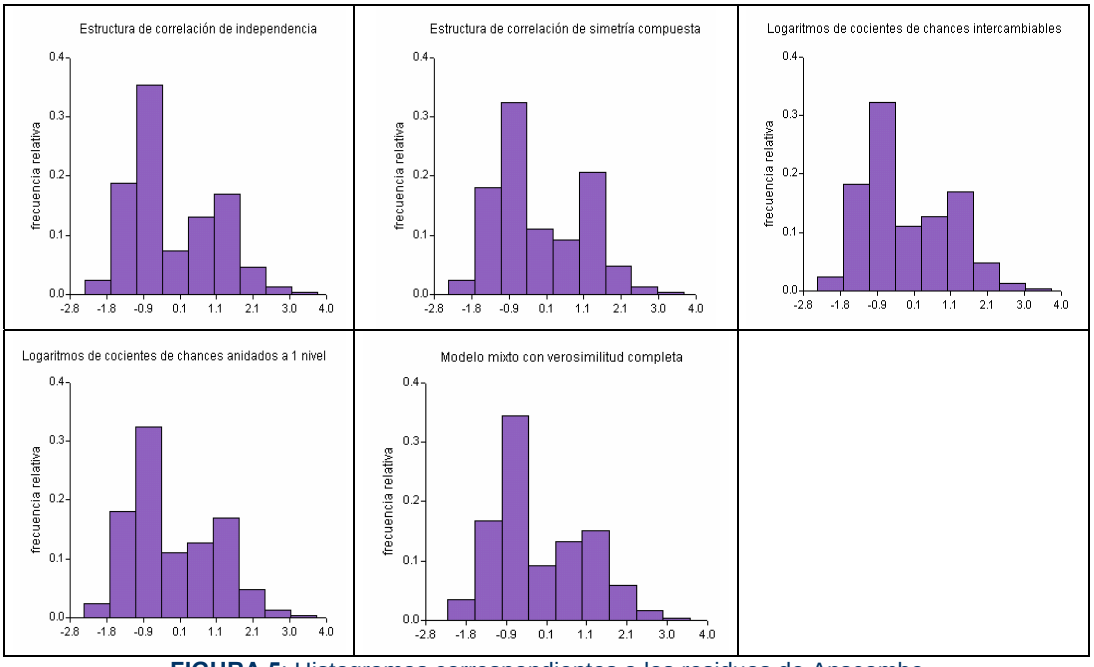

FIGURA 5: Histogramas correspondientes a los residuos de Anscombe para los modelos ajustados

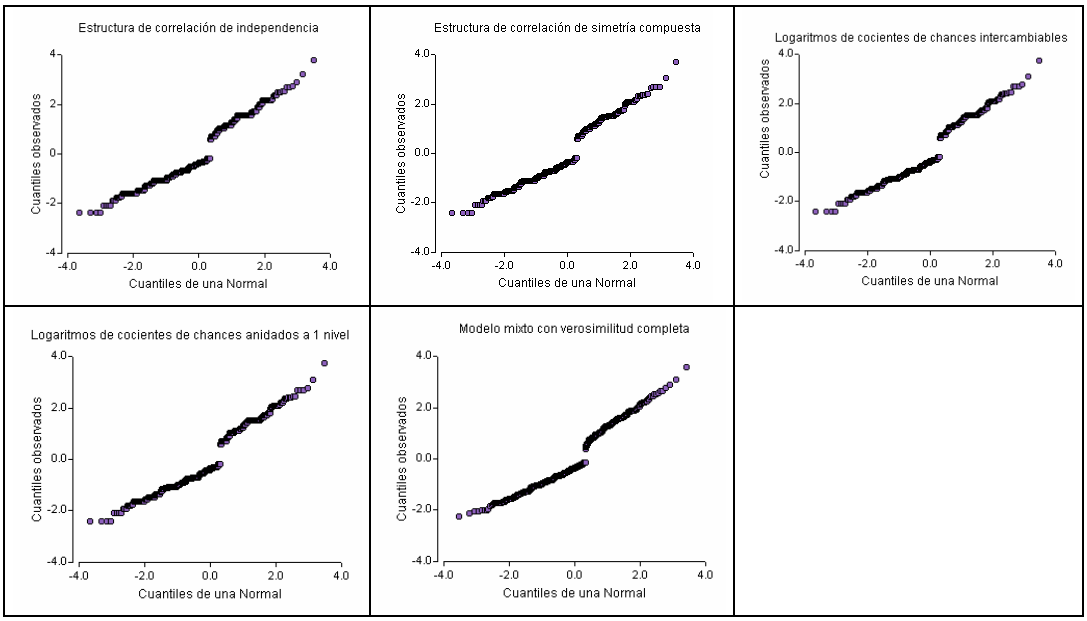

FIGURA 6: *Q-Q plots* correspondientes a los residuos de Anscombe para los modelos ajustados

Mientras que los histogramas muestran una distribución bimodal, señalando la existencia de dos poblaciones, los *Q-Q plots* exhiben una discontinuidad. Dicha discontinuidad no es en sí misma un patrón que indique el mal comportamiento de los residuos, como lo serían una marcada curvatura o un cambio brusco de pendiente, lo que no se verifica en los tramos analizados por separado.

Sin embargo, un análisis detallado revela que los residuos de Anscombe tienen signo positivo siempre que la variable respuesta vale 1 y signo negativo si la variable respuesta vale 0, al ser el primer término del numerador de la expresión [*73*] nulo en este caso. Como consecuencia, el estudio de la distribución de estos residuos no permite concluir acerca de la adecuación del modelo ni comparar a los distintos ajustes entre sí. Y, más aún, ningún otro tipo de residuo es útil a estos fines con observaciones binarias.

Al no seguir los residuos de Anscombe, en este caso, una distribución normal, es difícil detectar qué observaciones pueden calificarse de atípicas. No obstante, es interesante destacar aquéllas que poseen los mayores residuos en valor absoluto.

- **Observación 20**: corresponde a una mujer desocupada, con actitud no empresarial frente al desempleo, adversa al riesgo, con bajo nivel de creatividad y una visión

desfavorable hacia la actividad empresarial, que estudia Lic. en Administración en una universidad pública de la Ciudad Autónoma de Buenos Aires/Gran Buenos Aires. Esta alumna posee vocación emprendedora (VE=1) debido a que cuenta con una idea concreta de negocios, siendo la probabilidad estimada de 0.02 (TYPE=EXCH).

- **Observación 507**: corresponde a un hombre desocupado, sin actitud empresarial frente al desempleo, adverso al riesgo, con bajo nivel de creatividad y una visión desfavorable hacia la actividad empresarial, que estudia Ing. Industrial en una universidad privada de la Ciudad Autónoma de Buenos Aires/Gran Buenos Aires. Este alumno posee vocación emprendedora (VE=1) debido a que cuenta con una idea concreta de negocios, siendo la probabilidad estimada de 0.05 (TYPE=EXCH).
- **Observación 99**: corresponde a un hombre desocupado, sin actitud empresarial frente al desempleo, adverso al riesgo, con alto nivel de creatividad y una visión desfavorable hacia la actividad empresarial, que estudia Lic. en Economía en una universidad pública de la Ciudad Autónoma de Buenos Aires/Gran Buenos Aires. Este alumno posee vocación emprendedora (VE=1) debido a que cuenta con una idea concreta de negocios, siendo la probabilidad estimada de 0.09 (TYPE=EXCH).

Las tres observaciones referidas tienen residuos con signo positivo. Lo que explica su falta de ajuste es que los individuos carecen de la mayoría de los factores asociados con la presencia de vocación emprendedora, pero cuentan con una idea de negocios concreta que los clasifica como “con VE”. Evidentemente, otras características presentes en estos alumnos y no contempladas en el modelo, son las que explican su vocación emprendedora.

Un residuo alto indica que el modelo no logra ajustar la observación en cuestión, por lo que resulta informativo eliminarla y volver a ajustar el modelo con los datos restantes. Esto es equivalente a adicionar un parámetro para dicha observación, forzando un ajuste perfecto para la misma[34]. Dado que en este caso sólo 3 de 723 observaciones mostraron mal ajuste, los resultados prácticamente no se modifican al excluirlas, aunque las medidas de bondad del ajuste de los modelos mixtos se ven, naturalmente, levemente mejoradas.

7.5. Poder predictivo

Existen distintas formas de evaluar el poder predictivo de un modelo, varias de las cuales se aplican a continuación. En cada apartado se comentan las características de la medida utilizada y se presentan los resultados obtenidos para cada uno de los modelos ajustados.

Coeficiente de correlación

Un modo sencillo de evaluar el poder predictivo de un modelo consiste en calcular el coeficiente de correlación entre los valores observados y ajustados de la variable respuesta, para un mismo conjunto de datos. En una regresión logística, esta medida no necesariamente es no decreciente al tornarse el modelo más complejo y depende fuertemente del rango de valores de las covariables, por lo que resulta poco atractiva (Agresti, 2002).

En la Tabla 48 se presentan los coeficientes de correlación calculados para cada modelo. Si bien éstos prácticamente no difieren entre sí, la correlación bajo el modelo mixto con verosimilitud completa es ligeramente superior a la que exhiben los modelos marginales.

TABLA 48: Coeficiente de correlación $r\left(y_{ij}, \hat{\mu}_{ij}\right)$

Modelo ajustado	Correlación	Valor *p*
Independencia (TYPE=IND)	0.499	<0.001
Corr. intercambiable (TYPE=EXCH)	0.498	<0.001
Log OR intercambiables (LOGOR=EXCH)	0.498	<0.001
Log OR anidados a 1 nivel (LOGOR=NEST1)	0.498	<0.001
Verosimilitud completa (NLMIXED)	0.533	<0.001

[34] Una observación binaria en regresión logística suele tener menos influencia que una observación común en una regresión ordinaria, puesto que existe un límite para la distancia entre el valor observado y el ajustado (Agresti, 2002).

Tasa de error aparente

La tasa de error o de mala clasificación se define como la proporción de observaciones que un clasificador ubica, en promedio, incorrectamente en una clase. Una forma práctica de estimarla consiste en reclasificar las observaciones utilizando el modelo ajustado como clasificador y determinar qué porcentaje ha sido clasificado en forma incorrecta. La tasa de error se denomina aparente cuando es estimada con el mismo conjunto de datos con los que se ajusta el modelo. Este procedimiento puede subestimar la tasa de error de clasificación del modelo cuando es usado con observaciones futuras (Hand, 1996).

Dado el interés por indagar sobre el porcentaje de las observaciones que son correctamente clasificadas por el modelo, es necesario definir a partir de qué valor de probabilidad predicha se considera que el alumno posee vocación emprendedora (VE). Cuando no hay información a priori, es común que se adopte el valor 0.5 como punto de corte, lo cual supone que los costos de clasificación incorrecta son iguales en ambos sentidos. Sin embargo, al contar con la información brindada por los modelos estimados, es factible adoptar otro valor.

Las tablas de clasificación se confeccionan eligiendo el valor 0.4 como punto crítico de Pr(VE=1). Dicho valor corresponde al promedio, para los distintos modelos, de las probabilidades estimadas de que un alumno perteneciente a la categoría modal posea VE. De esta forma, la regla de decisión implica que si la probabilidad estimada es mayor o igual a 0.4 el alumno es clasificado en el grupo de alumnos “con VE”; caso contrario –i.e., si la probabilidad es menor a 0.4– se lo clasifica en el grupo de alumnos “sin VE”.

Con este criterio, se calculan probabilidades condicionales por fila: *(i)* dado que el alumno posee VE, la probabilidad de que el modelo lo clasifique “con VE” se denomina sensibilidad; *(ii)* dado que el alumno no posee VE, la probabilidad de que el modelo lo clasifique “sin VE” se llama especificidad. Los valores ubicados en la diagonal principal de la tabla de clasificación representan los porcentajes de observaciones correctamente clasificadas por el modelo y, los situados fuera de ella, representan las tasas de error aparente (Tabla 49).

TABLA 49: Denominación de las tasas de clasificación correcta e incorrecta

Valor observado	Pr(VE=1) ≥ 0.4	Pr(VE=1) < 0.4	Total
VE = 1	Sensibilidad	Error 1	100%
VE = 0	Error 2	Especificidad	100%

Seguidamente, se presentan las tablas de clasificación correspondientes a la regresión logística con observaciones independientes, al modelo marginal y al modelo mixto con verosimilitud completa. Si se ajusta un modelo de regresión logística ordinaria, el 62% de los alumnos con vocación emprendedora y el 80% de alumnos sin vocación emprendedora son correctamente clasificados (Tabla 50).

TABLA 50: Tasas de clasificación correspondientes al modelo de regresión logística ordinaria

Valor observado	Pr(VE=1) ≥ 0.4	Pr(VE=1) < 0.4	Total
VE = 1	**62.07%**	37.93%	100%
VE = 0	19.70%	**80.30%**	100%

Bajo el enfoque marginal, en la Tabla 51 se observa que los porcentajes de respuestas bien clasificadas con la estructura de independencia para la matriz de correlación de trabajo (TYPE=IND) coinciden con los obtenidos considerando que las observaciones no están correlacionadas. Sin embargo, si estos porcentajes se comparan con los correspondientes a las otras tres estructuras de dependencia, resulta que dicha opción otorga mayor especificidad pero menor sensibilidad. Es decir, si se incorpora explícitamente en el modelo la falta de independencia entre las observaciones, el modelo clasifica correctamente como poseedores de VE a un 16% más de alumnos que si se ignora la dependencia, pero aproximadamente un 13% menos de individuos sin VE son correctamente clasificados.

TABLA 51: Tasas de clasificación correspondientes al enfoque marginal

Estructura de dependencia	Valor observado	Pr(VE=1) ≥ 0.4	Pr(VE=1) < 0.4	Total
Independencia (TYPE=IND)	**VE = 1**	**62.07%**	37.93%	100%
	VE = 0	19.70%	**80.30%**	100%
Corr. intercambiable (TYPE=EXCH)	**VE = 1**	**78.16%**	21.84%	100%
	VE = 0	32.03%	**67.97%**	100%
Log de OR intercambiables (LOGOR=EXCH)	**VE = 1**	**78.16%**	21.84%	100%
	VE = 0	32.03%	**67.97%**	100%
Log OR anidados a 1 nivel (LOGOR=NEST1)	**VE = 1**	**78.16%**	21.84%	100%
	VE = 0	32.03%	**67.97%**	100%

En cuanto al porcentaje de respuestas correctamente clasificadas por el modelo mixto con verosimilitud completa, tal como surge de la Tabla 52, éste alcanza a un 70% de los alumnos con vocación emprendedora y a un 78% de alumnos sin vocación emprendedora.

TABLA 52: Tasas de clasificación correspondientes al modelo mixto con verosimilitud completa (NLMIXED)

Valor observado	Pr(VE=1) ≥ 0.4	Pr(VE=1) < 0.4	Total
VE = 1	**69.73%**	30.27%	100%
VE = 0	22.08%	**77.92%**	100%

Estos porcentajes difieren levemente de los reportados para los modelos marginales y, de la comparación entre los ajustes, surge que:

- Hay mayor sensibilidad y menor especificidad con el modelo mixto que considerando a las observaciones independientes entre sí.
- Hay mayor especificidad y menor sensibilidad con el modelo mixto que en los modelos marginales que especifican una estructura de dependencia.

En todos los casos, los porcentajes hallados resultan altamente satisfactorios. Más aún, si se tiene en cuenta que la variable que se modela está influida por múltiples factores psicológicos, sociales, económicos y culturales que no pueden ser incorporados en el modelo, y con sólo seis covariables se clasifica correctamente una alta proporción de las observaciones.

Curvas ROC

Si bien las tablas de clasificación pueden elaborarse para distintos puntos de corte, las curvas ROC *(Receiver Operating Characteristics)* son gráficos que sintetizan la relación entre la sensibilidad y uno menos la especificidad para todos los posibles puntos de corte, motivo por el cual resultan más 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.

El área debajo de la curva equivale a otra medida de poder predictivo llamada índice de concordancia *(c)*. Si se consideran todos los pares de observaciones *(r,s)* tales que $y_r=1$ e $y_s=0$, el estadístico *c* estima la probabilidad de que las predicciones y los resultados sean concordantes. Un valor de c=0.5 indica que las predicciones no son mejores que aciertos al azar –en tal caso, la curva ROC sería una línea recta– (Agresti, 2002).

Para los modelos estimados, se incluyen a continuación las curvas correspondientes a las distintas estructuras de dependencia elegidas bajo el enfoque marginal y al modelo mixto con verosimilitud completa (Figuras 7 a 9). Recuérdese que, en este aspecto, el modelo con estructura de independencia (TYPE=IND) equivale al modelo de regresión logística ordinaria.

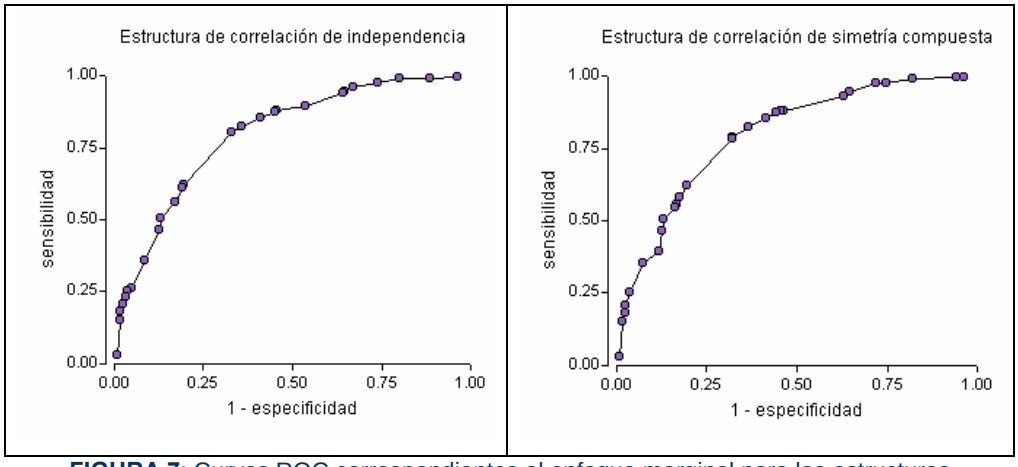

FIGURA 7: Curvas ROC correspondientes al enfoque marginal para las estructuras de independencia (TYPE=IND) y de simetría compuesta (TYPE=EXCH)

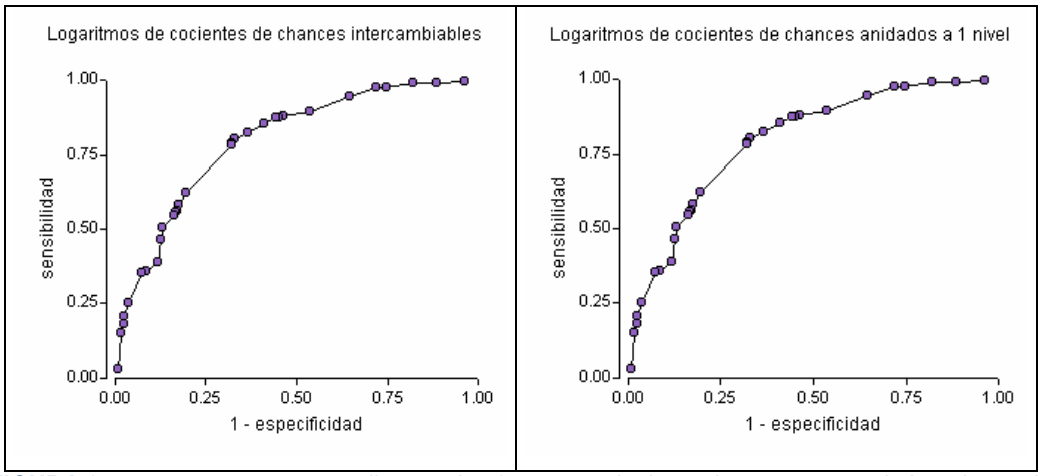

FIGURA 8: Curvas ROC correspondientes al enfoque marginal para las estructuras de logaritmos de cocientes de chances intercambiables (LOGOR=EXCH) y anidados a 1 nivel (LOGOR=NEST1)

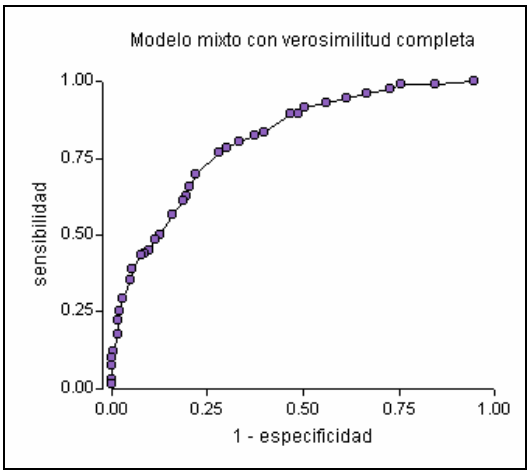

FIGURA 9: Curva ROC correspondiente al modelo mixto basado en la función de verosimilitud completa (NLMIXED)

La curva correspondiente al modelo mixto con verosimilitud completa es ligeramente más alta que las demás. Si se observa el gráfico atentamente, puede apreciarse que para un nivel de especificidad de 0.75, la mayor sensibilidad se logra precisamente con dicho modelo. Aunque esto puede considerarse como una ventaja, debe tenerse en cuenta que las curvas son estimadas, por lo cual las proporciones están sujetas a error y las diferencias observadas en el gráfico pueden no ser estadísticamente significativas.

Una forma aproximada de calcular el área debajo de las curvas ROC es sumando las áreas de trapecios conformados entre dos puntos de la curva y el eje de abscisas. Construyendo una curva discreta con 100 puntos de corte, los valores obtenidos se incluyen en la Tabla 53.

TABLA 53: Área debajo de la curva ROC para los distintos ajustes

Modelo ajustado	Área
Independencia (TYPE=IND)	0.7975
Corr. intercambiable (TYPE=EXCH)	0.7966
Log OR intercambiables (LOGOR=EXCH)	0.7966
Log OR anidados a 1 nivel (LOGOR=NEST1)	0.7967
Verosimilitud completa (NLMIXED)	0.8158

Como puede apreciarse, las áreas calculadas para los modelos marginales son idénticas entre sí, cualquiera sea la estructura de dependencia elegida. En consonancia con lo mencionado anteriormente, el área bajo la curva para el modelo mixto con verosimilitud completa es apenas superior. No obstante, la diferencia es mínima como para afirmar que la predictibilidad, evaluada de este modo, difiere entre ambos enfoques. Lo que es importante destacar es que el valor hallado confiere a los modelos un alto poder predictivo.

Tasa de error por validación cruzada "leave-one-out"

Otra alternativa para evaluar el poder predictivo del modelo consiste en utilizar los métodos de validación cruzada. Éstos consisten en generar un clasificador con distintos subconjuntos de datos y juzgar su funcionamiento usando los datos restantes, repitiendo el procedimiento para distintos subconjuntos y promediando los resultados. Una de sus variantes, llamada *leave-one-out*, aprovecha una única observación para probar el clasificador construido en base a las restantes *(n–1)* observaciones y repite el proceso *n* veces[35]. Este enfoque posee la ventaja de que, en cada caso, el tamaño del conjunto de datos se mantiene igual, lo que garantiza que la tasa de error estimada sea aproximadamente insesgada. El método es computacionalmente intensivo y hay evidencia de que posee una varianza relativamente grande en pequeñas muestras (Hand, 1996).

El método *leave-one-out* difiere del denominado *Jackknife*, a pesar de ser superficialmente similar al anterior. Aún cuando ambos omiten una observación por ciclo, la validación cruzada se utiliza para estimar la tasa de error y *Jackknife* para estimar el sesgo de un estadístico. En tal caso, se computa el estadístico de interés con cada subconjunto de datos y el promedio de los valores estimados se compara con el valor hallado con la muestra completa; también se pueden obtener estimadores *Jackknife* del error estándar. Este método puede aplicarse para estimar el sesgo de la tasa de error, pero el procedimiento resulta mucho más complicado que la validación *leave-one-out* (comp.ai.neural-nets, 2000).

Para cuantificar la predictibilidad de cada uno de los modelos ajustados, se aplica el método de validación cruzada estimando los parámetros de regresión con 722 datos, clasificando a la observación excluida como "con VE" si su probabilidad asignada es mayor o igual a 0.4 o "sin VE" si dicha probabilidad es menor a 0.4 y repitiendo el procedimiento 723 veces. Para ello se corre una macro de SAS que se incluye en el Anexo D.

Bajo el enfoque marginal, las tasas de error por validación cruzada (Tabla 54) son idénticas a las tasas de error aparente (Tabla 51). Ello indica que aún cuando las probabilidades estimadas al utilizar la totalidad de la muestra difieren de las predicciones obtenidas por validación cruzada, ninguna observación resulta clasificada en un grupo distinto al que inicialmente pertenecía.

TABLA 54: Tasas de clasificación por validación cruzada correspondientes al enfoque marginal

Estructura de dependencia	Valor observado	Pr(VE=1) ≥ 0.4	Pr(VE=1) < 0.4	Total
Independencia (TYPE=IND)	**VE = 1**	**62.07%**	37.93%	100%
	VE = 0	19.91%	**80.09%**	100%
Corr. intercambiable (TYPE=EXCH)	**VE = 1**	**78.16%**	21.84%	100%
	VE = 0	32.03%	**67.97%**	100%
Log OR intercambiables (LOGOR=EXCH)	**VE = 1**	**78.16%**	21.84%	100%
	VE = 0	32.03%	**67.97%**	100%
Log OR anidados a 1 nivel (LOGOR=NEST1)	**VE = 1**	**78.16%**	21.84%	100%
	VE = 0	32.03%	**67.97%**	100%

Si bien la predictibilidad del modelo mixto también puede considerarse alta, las tasas de error por validación cruzada (Tabla 55) son algo superiores a las tasas de error aparente (Tabla 52) –i.e., menor especificidad y sensibilidad–. Ello podría explicarse por la mayor variabilidad, debida al distinto alcance que tiene la inferencia bajo este enfoque, y porque el procedimiento no respeta el tamaño relativo de los *clusters*, lo cual puede incidir sobre la estimación de los efectos aleatorios.

TABLA 55: Tasas de clasificación por validación cruzada correspondientes al modelo mixto con verosimilitud completa (NLMIXED)

Valor observado	Pr(VE=1) ≥ 0.4	Pr(VE=1) < 0.4	Total
VE = 1	**65.52%**	34.48%	100%
VE = 0	24.03%	**75.97%**	100%

[35] Otras variantes son: *leave-v-out*, donde se utilizan *v* observaciones por ciclo para evaluar el clasificador, y *k-fold*, en la

Con respecto a los valores predichos con el método *leave-one-out*, éstos tienden a ser levemente superiores a los obtenidos con la muestra original (Tabla 56). Para el modelo mixto con verosimilitud completa se observa que la diferencia promedio entre ambas estimaciones es mayor, i.e., las predicciones con la muestra por validación cruzada distan más de los valores predichos con la muestra original.

TABLA 56: Estadísticos descriptivos de la diferencia entre los valores predichos por validación cruzada y con la muestra original

Modelo ajustado	% valores predichos por CV mayores a los valores predichos con la muestra original	Media	Error estándar	Mínimo	Máximo
Independencia (TYPE=IND)	63.90%	0.00002	0.005	–0.017	0.012
Corr. intercambiable (TYPE=EXCH)	62.38%	0.00003	0.005	–0.018	0.016
Log OR intercambiables (LOGOR=EXCH)	63.90%	0.00004	0.005	–0.017	0.015
Log OR anidados a 1 nivel (LOGOR=NEST1)	63.90%	0.00003	0.005	–0.017	0.014
Verosimilitud completa (NLMIXED)	63.90%	0.00033	0.012	–0.080	0.039

Los resultados hallados permiten concluir que los modelos estimados poseen un elevado poder predictivo. Además, el hecho de que ningún valor predicho bajo el enfoque marginal sea clasificado en un grupo distinto al que había sido asignado al trabajar con la muestra completa, implica que los coeficientes son estables y que no existen datos puntuales que sean influyentes. En cuanto al modelo mixto con verosimilitud completa, si bien la predictibilidad puede considerarse alta, es el que menos satisfactoriamente se comporta.

7.6. Interpretación de coeficientes

Los parámetros estimados bajo las distintas estrategias difieren en su interpretación, aún cuando todos ellos se traducen en cocientes de chances condicionales. 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,

cual se definen *k* particiones mutuamente excluyentes que alternativamente se usan para evaluar la regla de clasificación

aunque condicionan sobre las otras covariables del modelo (Tabla 57). Si los parámetros son estimados mediante un modelo mixto, los mismos representan cocientes de chances condicionales controlando por el efecto de *cluster*, i.e., correspondientes a una facultad determinada (Tabla 58).

La interpretación se ejemplifica con los resultados correspondientes al modelo marginal con la estructura de correlación de simetría compuesta (TYPE=EXCH) y al modelo mixto con verosimilitud completa (NLMIXED).

TABLA 57: Coeficientes estimados bajo el enfoque 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 se multiplican por un factor de 4.5 si visualiza favorablemente la actividad 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 alta creatividad.

TABLA 58: Coeficientes estimados bajo el enfoque mixto con verosimilitud completa

Covariable	exp(β)	Interpretación
GENERO	2.68	Controlando por las demás covariables y para una facultad determinada, las chances de que 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 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 posea vocación emprendedora se multiplican por un factor de 3 si tiene actitud emprendedora frente al desempleo.
VISION	4.94	Controlando por las demás covariables y para una facultad determinada, las chances de que 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 posea vocación emprendedora se multiplican por un factor de 2.2 si su propensión al riesgo es alta.
CREATIV	1.73	Controlando por las demás covariables y para una facultad determinada, las chances de que posea vocación emprendedora se multiplican por un factor de 1.7 si tiene alta creatividad.

construida por las restantes (Hand, 1996; Bao, 2000).

El intercepto, no incluido en las tablas anteriores, consiste en un estimador de las chances –i.e., cociente entre la frecuencia de individuos con VE=1 y VE=0– para la subpoblación en la que todas las covariables son iguales a cero. El mismo se obtiene promediando a través de los *clusters* bajo el enfoque marginal, o para un *cluster* típico en los modelos mixtos.

La principal distinción entre la inferencia realizada con el enfoque marginal y con el enfoque mixto es que los coeficientes de regresión describen la respuesta poblacional promedio o la respuesta específica de cada *cluster* ante un cambio en el nivel de las covariables. Por tal motivo, pueden responderse distintas preguntas de investigación de acuerdo al enfoque elegido. Mediante el modelo marginal es posible responder, e.g.:

- Si la proporción de alumnos con vocación emprendedora es igual en la subpoblación de mujeres y de hombres.
- Si la asociación entre vocación emprendedora y género cambia según la situación ocupacional del alumno.

Las preguntas anteriores también pueden responderse con el enfoque mixto, aunque en este caso la asociación se analizaría, e.g., para hombres y mujeres del mismo *cluster*. Asimismo, aplicando el modelo mixto podría responderse:

- Si la probabilidad de que un alumno posea vocación emprendedora es mayor en la U1 que en la U4.
- Si el efecto del género sobre la probabilidad de presencia de vocación emprendedora es mayor en la U1 que en la U4.

Ello no sería posible bajo el enfoque marginal, debido a que éste no estima los coeficientes que corresponden a cada unidad de muestreo, sino promedios poblacionales. Por consiguiente, las preguntas y objetivos que guían la investigación actúan como determinantes del enfoque estadístico a elegir, independientemente de otras consideraciones.

7.7. Resumen de resultados

Con el enlace logístico y las covariables GENERO, OCUPADO, ACTITUD, VISION, RIESGO y CREATIV, se han ajustado modelos marginales, mixtos y de regresión logística ordinaria para estimar la probabilidad de presencia de vocación emprendedora en alumnos universitarios. Las covariables se encuentran altamente asociadas con la variable respuesta pero no entre sí, lo cual excluye la presencia de multicolinealidad. Aún cuando se ha probado la adición de interacciones dobles en el predictor lineal, el modelo de efectos principales aparece como adecuado.

A pesar de la baja correlación intra-*cluster*, el cociente de chances común estimado con la estructura de logaritmos de cocientes de chances intercambiables es estadísticamente significativo. Utilizando la función de verosimilitud completa, el estimador de la raíz cuadrada de la varianza de la distribución de los efectos aleatorios también es significativamente distinto de cero. Ello valida el supuesto de que existe asociación entre las observaciones de una misma facultad.

Los resultados indican que la probabilidad de que un alumno posea vocación emprendedora se incrementa si el sujeto es hombre, se encuentra ocupado, tiene una actitud emprendedora frente al desempleo, valora positivamente la actividad empresarial, es propenso al riesgo o posee un alto nivel de creatividad. Teniendo en cuenta la tabla de contingencia parcial con mayor frecuencia, se define una categoría modal. Para un individuo perteneciente a la misma, la probabilidad se estima en 0.409, 0.421 y 0.387, según el enfoque adoptado –i.e., marginal, mixto con verosimilitud completa o regresión logística ordinaria–.

Al comparar la eficiencia entre los distintos métodos empleados, resulta que:

- Los intervalos de confianza para los parámetros son más estrechos si se modela la correlación entre las observaciones con la estructura de simetría compuesta. Entre los modelos mixtos, el basado en la verosimilitud condicional provee estimaciones más precisas que el modelo con verosimilitud completa, aunque no permite obtener una estimación directa del intercepto al condicionarlo fuera del modelo.

- Los errores estándares estimados presentan escasa variabilidad entre los distintos modelos ajustados bajo el enfoque marginal –lo que denota que la estimación es robusta– y entre ambos modelos mixtos, pero alta variabilidad si se comparan estos enfoques entre sí o con respecto a una regresión logística ordinaria.

Si se analiza la predictibilidad del modelo mediante el cálculo de la tasa de error aparente, se observa que, dependiendo del ajuste realizado, no más del 32% de los individuos con VE=0 y del 38% de quienes tienen VE=1 son mal clasificados. Comparativamente:

- La mayor sensibilidad se alcanza con los modelos marginales con estructuras de correlación de simetría compuesta (TYPE=EXCH) o de logaritmos de cocientes de chances intercambiables o anidados a un nivel (LOGOR=EXCH y LOGOR=NEST1).
- La mayor especificidad se logra con el modelo de regresión logística ordinaria y con el modelo mixto con verosimilitud completa (NLMIXED).

Empleando las curvas ROC, no se aprecian diferencias entre los distintos ajustes, exhibiendo todos los modelos un alto poder predictivo al calcularse el área debajo de la curva en 0.80. En cuanto a las tasas de error por validación cruzada *leave-one-out*, éstas son idénticas a las tasas de error aparente para los modelos marginales. Para el modelo mixto con verosimilitud completa, son ligeramente superiores.

Antes de comparar los métodos de inferencia basada en el diseño muestral y basada en modelos, en la Tabla 59 se sintetizan algunos de los resultados hallados para tres de los ajustes: modelo marginal con estructura de correlación de simetría compuesta (TYPE=EXCH), modelo mixto con verosimilitud completa (NLMIXED) y modelo de regresión logística ordinaria.

TABLA 59: Resultados comparados para distintos modelos ajustados

	Modelo marginal	**Modelo mixto**	**Regresión logística ordinaria**
Estimación	Se modela la estructura de dependencia entre las observaciones de un mismo *cluster.*	Se incorpora un efecto aleatorio para cada *cluster* en el predictor lineal.	Se considera que las observaciones son independientes.
	Todas las covariables son estadísticamente significativas.	Todas las covariables son estadísticamente significativas.	Todas las covariables son estadísticamente significativas.
Inferencia	Amplitud relativa de los IC: 0.89	Amplitud relativa de los IC: 0.98	Amplitud relativa de los IC: 0.93
	Pr(y_{ij}=1) para un individuo de la categoría modal: $\hat{\mu}_{..} = 0.41$ $IC = (0.32, 0.50)$	*Pr(yij=1/U_i=0)* para un individuo de la categoría modal: $\hat{\mu}_{..}/U_i = 0.42$ $IC = (0.29, 0.55)$	*Pr(y_{ij}=1)* para un individuo de la categoría modal: $\hat{\mu}_{..} = 0.39$ $IC = (0.31, 0.47)$
	Máximo valor predicho: 0.88	Máximo valor predicho: 0.94	Máximo valor predicho: 0.87
Predictibilidad del modelo	■ Tasa de error aparente: Error 1 = 21.84% Error 2 = 32.03% ■ Tasa de error CV: Error 1 = 21.84% Error 2 = 32.03%	■ Tasa de error aparente: Error 1 = 30.27% Error 2 = 22.08% ■ Tasa de error CV: Error 1 = 34.48% Error 2 = 24.03%	■ Tasa de error aparente: Error 1 = 37.93% Error 2 = 19.70% ■ Tasa de error CV: Error 1 = 37.93% Error 2 = 19.91%
Interpretación de los coeficientes	Cocientes de chances promedio para la población *(population average inference)*, promediados a través de los *clusters*.	Cocientes de chances condicionales al *cluster (cluster specific inference)*.	Cocientes de chances promedio para la población *(population average inference)*, ignorando la estructura de *clusters*.

8. COMPARACIÓN ENTRE MÉTODOS DE INFERENCIA

Luego de realizar inferencia clásica (sección 6) e inferencia basada en modelos (sección 7), resta comparar los resultados hallados por ambas vías. En la sección 6 se estimó en **0.40** la proporción de alumnos universitarios con vocación emprendedora (VE), calculando el total de alumnos con VE sobre el total de alumnos encuestados, ponderado por el tamaño de cada facultad. Este valor es muy próximo al obtenido mediante los modelos lineales generalizados marginales (**0.406**), mixtos (**0.421**) y de regresión logística ordinaria (**0.387**), al calcular la probabilidad de que un alumno perteneciente a la denominada categoría modal posea vocación emprendedora. Sin embargo, los modelos incorporan información acerca de las covariables que es completamente ignorada al hacer inferencia clásica.

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, i.e., 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. Por tal motivo, con la inferencia clásica no puede estimarse la proporción de alumnos pertenecientes a la categoría modal que poseen vocación emprendedora –i.e., no puede estimarse la proporción de alumnos hombres, ocupados, con actitud no emprendedora frente al desempleo, que visualizan favorablemente la actividad emprendedora, adversos al riesgo y con baja creatividad con VE–, por no disponer de la información requerida.

A continuación, se estima la proporción de alumnos universitarios de género masculino y femenino con vocación emprendedora bajo ambos métodos, calculándose los intervalos de confianza asociados. Como ya se demostrara, el género es una variable relacionada con la vocación emprendedora:

- Mediante la inferencia clásica (sección 6), la proporción de hombres con VE es significativamente mayor que la proporción de mujeres con VE.
- Con la inferencia basada en modelos (sección 7), la probabilidad –función uno a uno de las chances– de poseer VE es mayor si el alumno es hombre que si se trata de una mujer.

Para poder obtener resultados comparables mediante el uso de modelos, a GENERO 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.

Las probabilidades se estiman utilizando los coeficientes de los modelos: *(i)* marginal con estructura de correlación de simetría compuesta (TYPE=EXCH), *(ii)* mixto con verosimilitud completa (NLMIXED) y *(iii)* de regresión logística ordinaria. Los valores asignados a las covariables, representativos de los porcentajes muestrales, se presentan en la Tabla 60. A continuación se observa el estimador de la probabilidad para cada modelo, aplicando la función de enlace inversa:

$$\hat{\mu}_{ij(MARGINAL)} = \frac{e^{(-3.804+0.930*GENERO+1.003*OCUPADO+1.040*ACTITUD+1.503*VISION+0.766*RIESGO+0.520*CREATIV)}}{1+e^{(-3.804+0.930*GENERO+1.003*OCUPADO+1.040*ACTITUD+1.503*VISION+0.766*RIESGO+0.520*CREATIV)}}.$$

$$\hat{\mu}_{ij}\,/\,U_{i(MIXTO)} = \frac{e^{(-3.937+0.985*GENERO+1.035*OCUPADO+1.085*ACTITUD+1.598*VISION+0.787*RIESGO+0.547*CREATIV+U_i)}}{1+e^{(-3.937+0.985*GENERO+1.035*OCUPADO+1.085*ACTITUD+1.598*VISION+0.787*RIESGO+0.547*CREATIV+U_i)}}.$$

$$\hat{\mu}_{ij(RLO)} = \frac{e^{(-3.919+0.801*GENERO+1.056*OCUPADO+1.074*ACTITUD+1.601*VISION+0.818*RIESGO+0.484*CREATIV)}}{1+e^{(-3.919+0.801*GENERO+1.056*OCUPADO+1.074*ACTITUD+1.601*VISION+0.818*RIESGO+0.484*CREATIV)}}.$$

TABLA 60: Valores asignados a las covariables en los modelos, correspondientes a las proporciones muestrales

Hombres	Mujeres
GENERO = 1 OCUPADO = 0.5116 ACTITUD = 0.4211 VISION = 0.7979 RIESGO = 0.3284 CREATIV = 0.2358	GENERO = 0 OCUPADO = 0.6008 ACTITUD = 0.3427 VISION = 0.6855 RIESGO = 0.2298 CREATIV = 0.3387

Una primera diferencia a resaltar entre ambos métodos es que al efectuar inferencia clásica, se emplean 475 observaciones para la subpoblación de alumnos de género masculino y 248 observaciones para la subpoblación de alumnos de género femenino (ver Tabla 7). Sin embargo, con los modelos se estiman las proporciones utilizando la totalidad de los datos (723) cualquiera sea el valor de las covariables. En la Tabla 61 se indica la estimación puntual y la amplitud de los intervalos de confianza asociados.

TABLA 61: Proporciones estimadas por género bajo ambos métodos de inferencia

Subpoblación	Inferencia	Proporción estimada	Amplitud del intervalo de confianza
Hombres con VE=1	Inferencia Clásica	0.469	0.137
	Enfoque marginal Corr. intercambiable (TYPE=EXCH)	0.414	0.143
	Enfoque mixto (NLMIXED) U_i=0	0.424	0.226
	Enfoque mixto (NLMIXED) U_i promedio	0.392	0.220
	Regresión logística ordinaria	0.386	0.102
Mujeres con VE=1	Inferencia Clásica	0.316	0.130
	Enfoque marginal Corr. intercambiable (TYPE=EXCH)	0.188	0.107
	Enfoque mixto (NLMIXED) U_i=0	0.184	0.169
	Enfoque mixto (NLMIXED) U_i promedio	0.165	0.154
	Regresión logística ordinaria	0.187	0.106

En la subpoblación de hombres es prácticamente igual de eficiente aplicar inferencia clásica o el enfoque marginal, y más eficiente aún si se opta por el modelo que supone observaciones independientes. Para la subpoblación de mujeres, el modelo marginal y el de regresión logística ordinaria brindan 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 halladas 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 –i.e., efecto aleatorio nulo–, se adiciona al predictor lineal el promedio ponderado de los efectos aleatorios. Tal como se había comentado anteriormente, 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 –i.e., "controlando" por la facultad a la que asiste el alumno–, éste es el único método adecuado a pesar de que la inferencia sea menos precisa. La amplitud de los intervalos de confianza, para los distintos ajustes, puede verse graficada en la Figura 10.

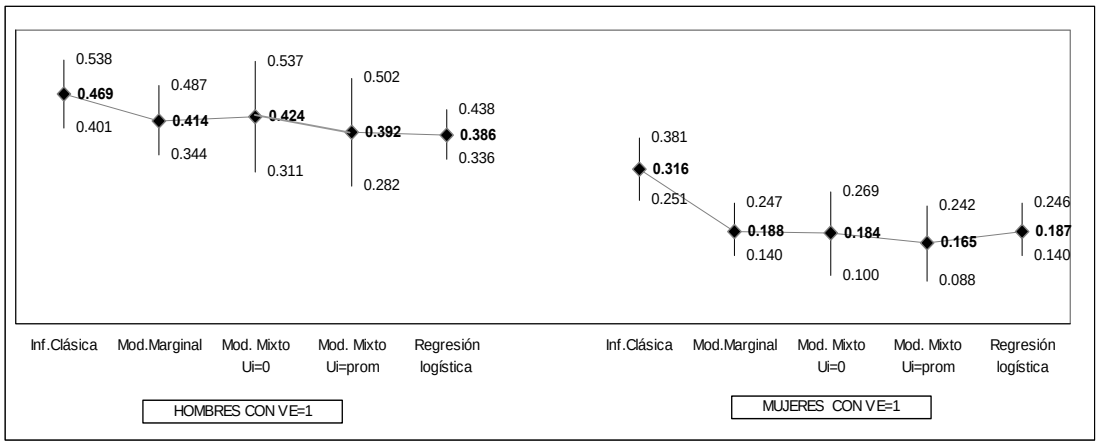

FIGURA 10: Estimación puntual y límites del intervalo de confianza para la proporción de hombres y mujeres con VE mediante ambos métodos de inferencia

Dado que la ventaja de los modelos mixtos es poder efectuar inferencia específica para cada *cluster*, es interesante ver cómo se estiman las proporciones por género, ejemplificando con la U1 por ser la de mayor tamaño. Para ello se calculan los porcentajes muestrales correspondientes a las covariables distintas de GENERO para hombres y mujeres de la U1, adicionándole al predictor lineal el efecto aleatorio estimado para dicha facultad (Tabla 62). En este caso, la interpretación ya no es general, sino específica para la U1:

> La probabilidad estimada de que un alumno que asiste a la U1 posea VE es de 0.51 si es hombre, mientras que dicha probabilidad es de 0.29 si se trata de una mujer.

TABLA 62: Proporciones estimadas con el modelo mixto con verosimilitud completa (NLMIXED) para la U1

Subpoblación	Proporción estimada	Amplitud del intervalo de confianza
Hombres con VE=1	0.509	0.229
Mujeres con VE=1	0.291	0.224

Una segunda apreciación que surge de la Tabla 61, 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. Si en lugar del estimador presentado en [*12*] se calculara la proporción muestral como alumnos con VE sobre el total de alumnos, para cada género, las proporciones serían de 0.25 para las mujeres y de 0.42 para los hombres. Es decir, más próximas a las estimadas con los modelos.

La diferencia se explica porque la fórmula utilizada le otorga más peso a los *clusters* de mayor tamaño. Esto se ejemplifica en la Tabla 63, comprobando que las proporciones estimadas disminuyen si se excluye a la U1 que es la que tiene más alumnos. Por consiguiente, la estimación mediante el uso de modelos depende menos del tamaño relativo de los *clusters* que la proporción estimada con inferencia clásica.

TABLA 63: Proporciones estimadas bajo la inferencia clásica con y sin la U1

Subpoblación	Observaciones ajustadas	Proporción estimada	Amplitud del intervalo de confianza
Hombres con VE=1	Incluyendo la U1	0.469	0.137
	Excluyendo la U1	0.420	0.153
Mujeres con VE=1	Incluyendo la U1	0.316	0.130
	Excluyendo la U1	0.243	0.195

Esta es una afirmación general en modelos estadísticos: las estimaciones basadas en modelos usan los estimadores de los parámetros del mismo para estimar valores esperados, o funciones de valores esperados. En la inferencia no basada en modelos se usan directamente promedios ponderados por los tamaños de grupos observados, *clusters* en este caso. En áreas relacionadas, e.g. modelos lineales desbalanceados, el uso de estimaciones basadas en modelos se considera más informativo que el uso de estimaciones directas de medias (Milliken & Johnson, 1992).

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, ya que las instituciones no procesan dicha información, se propone aplicar las proporciones muestrales de OCUPADO=1 y OCUPADO=0 por carrera al total de individuos de cada *cluster*, las que resultan:

- Hombres que estudian economía y administración: 57% ocupados.
- Hombres que estudian ingeniería: 47% ocupados.
- Mujeres que estudian economía y administración: 59% ocupadas.
- Mujeres que estudian ingeniería: 65% ocupadas.

Las medias estimadas por género y situación ocupacional se presentan en la Tabla 64.

TABLA 64: Proporciones estimadas por género y situación ocupacional bajo ambos métodos de inferencia

Subpoblación	Inferencia	Proporción estimada	Amplitud del intervalo de confianza
Hombres ocupados con VE=1	Inferencia clásica	0.590	0.204
	Enfoque marginal Corr. intercambiable (TYPE=EXCH)	0.535	0.208
	Enfoque mixto (NLMIXED) U_i=0	0.550	0.240
	Enfoque mixto (NLMIXED) U_i promedio	0.519	0.242
	Regresión logística ordinaria	0.513	0.134
Mujeres ocupadas con VE=1	Inferencia clásica	0.399	0.228
	Enfoque marginal Corr. intercambiable (TYPE=EXCH)	0.257	0.164
	Enfoque mixto (NLMIXED) U_i=0	0.256	0.214
	Enfoque mixto (NLMIXED) U_i promedio	0.232	0.200
	Regresión logística ordinaria	0.260	0.137
Hombres desocupados con VE=1	Inferencia clásica	0.341	0.132
	Enfoque marginal Corr. intercambiable (TYPE=EXCH)	0.297	0.132
	Enfoque mixto (NLMIXED) U_i=0	0.303	0.220
	Enfoque mixto (NLMIXED) U_i promedio	0.277	0.208
	Regresión logística ordinaria	0.268	0.116
Mujeres desocupadas con VE=1	Inferencia clásica	0.137	0.121
	Enfoque marginal Corr. intercambiable (TYPE=EXCH)	0.112	0.079
	Enfoque mixto (NLMIXED) U_i=0	0.097	0.112
	Enfoque mixto (NLMIXED) U_i promedio	0.079	0.092
	Regresión logística ordinaria	0.109	0.085

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*– vale 1

o 0. Esto es así porque, en tales casos, no puede estimarse la varianza al hacerse nulo el denominador de la sumatoria que aparece en el segundo término de la expresión [*13*]. Los *clusters* excluidos en cada grupo para el cálculo, son:

- Hombres ocupados → U8.
- Mujeres ocupadas → U3, U10 y U12.
- Hombres desocupados → U3.
- Mujeres desocupadas → U7, U8, U9, U10 y U11.

A partir de la Tabla 64 resulta que:

- Los intervalos de confianza para la inferencia clásica no son en ningún caso los más estrechos, hallándose siempre por encima de los asociados a los modelos de regresión logística ordinaria y marginales. Incluso para el grupo “mujeres ocupadas”, esta inferencia es más imprecisa que la correspondiente al modelo mixto.
- A excepción de la combinación “hombres ocupados”, en donde no se aprecian diferencias, los intervalos de confianza se achican si en el modelo de verosimilitud completa se considera el efecto de *cluster* promedio. I.e., con un efecto aleatorio promedio la inferencia es tanto o más precisa que si se hace en referencia a un *cluster* típico.
- Los intervalos de confianza son más estrechos si se aplica inferencia basada en modelos de regresión logística ordinaria o marginales. Para la combinación “mujeres desocupadas”, el modelo marginal con estructura de correlación de simetría compuesta ofrece la estimación más precisa.
- Respecto de las estimaciones puntuales, aquéllas que surgen de aplicar la inferencia clásica son las más altas, denotando el impacto de la ponderación por el tamaño del *cluster*.

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. Inicialmente, al considerar sólo la variable GENERO, los intervalos de confianza no diferían demasiado de los correspondientes a los modelos. Sin

embargo, 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.

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 –en particular, acerca del tamaño total del *cluster*–.
- 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. Esto representa un inconveniente, dado que dicha información puede no estar disponible: *(a)* por fallas en los marcos muestrales; *(b)* 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. Dicha configuración depende de que exista información previa sobre las covariables seleccionadas.

- 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 las respectivas tablas de contingencia parciales sólo contienen celdas vacías.

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 –e.g., 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.

9. CONCLUSIONES

Si se desean modelar respuestas binarias captadas mediante un muestreo por conglomerados en dos etapas, es necesario considerar la dependencia entre las observaciones pertenecientes a un mismo *cluster*. Ello puede efectuarse: *(a)* especificando una estructura de correlación o modelando los logaritmos de los cocientes de chances; *(b)* en forma inducida, a través de la inclusión de efectos aleatorios en el predictor lineal. Alternativamente, pueden formularse tres modelos para describir las respuestas correlacionadas en función de un conjunto de covariables:

- **Modelo marginal**: se modelan, por separado, la esperanza marginal de la variable respuesta como función de las covariables y la estructura de dependencia entre las observaciones pertenecientes a un mismo *cluster*.
- **Modelo mixto con verosimilitud completa**: se incluyen en el predictor lineal efectos fijos y aleatorios, siguiendo estos últimos una distribución paramétrica –que en general se propone normal– a través de la población. Dado el *cluster*, las observaciones se consideran independientes entre sí y con densidad en la familia exponencial. Este modelo combina la información provista por las comparaciones entre y dentro de los *clusters*.
- **Modelo mixto con verosimilitud condicional**: el predictor lineal contiene solamente a los efectos fijos, los que se estiman condicionando sobre los efectos aleatorios. En este caso, no es necesario especificar una distribución de probabilidad para estos últimos. Este modelo sólo aprovecha la variabilidad intra-*cluster* para estimar los parámetros.

Bajo el **enfoque marginal**, los parámetros se estiman por el método de ecuaciones de estimación generalizadas *(GEE)*. Aún si la correlación entre pares de respuestas es baja, los errores estándares estimados mediante *GEE* son más robustos que los que surgen de un modelo ordinario de regresión logística. Ello es así porque el estimador *"sandwich"* de la varianza de los parámetros incorpora información acerca de la dependencia muestral que

exhiben los datos. Este enfoque es más fácil de implementar y es computacionalmente menos intensivo que el mixto.

Entre las distintas **estructuras de dependencia** disponibles para los modelos marginales, el supuesto de equi-correlación resulta preferible si la correlación intra-*cluster* es baja. Esta opción ofrece una buena predictibilidad combinada con estimaciones eficientes de los parámetros, a la vez que incurre en el costo de estimar un único coeficiente adicional correspondiente a la correlación común.

Bajo el **enfoque mixto**, la formulación más simple es la de intercepto aleatorio: ella supone que el efecto aleatorio estimado se adiciona al intercepto del predictor lineal. Con el modelo de verosimilitud completa, ésta es una cantidad por la cual todas las mediciones del *cluster* se ven incrementadas o disminuidas respecto de un *cluster* típico cuyo efecto aleatorio sea nulo. Ajustando el modelo con verosimilitud condicional, el intercepto se condiciona fuera del modelo y permanece desconocido. No obstante, aprovechando la relación que existe con la función de verosimilitud parcial propia del análisis de sobrevida, puede obtenerse una aproximación al valor del intercepto a partir de la estimación de la función de sobrevida.

El modelo de **regresión logística ordinaria** comúnmente se emplea en lugar de los modelos antes comentados, por ser más sencillo. Éste incluye sólo efectos fijos en el predictor lineal y trata a las observaciones como si fueran independientes, ignorando la estructura de *clusters*. Si la correlación intra-*cluster* es baja, las estimaciones obtenidas son semejantes y la inferencia puede resultar aún más precisa que bajo los enfoques marginal y mixto. Sin embargo, no debe olvidarse que los errores estándares estimados pueden ser inconsistentes si se supone, en forma errónea, que no existe dependencia entre las observaciones.

Si se opta por el enlace *logit*, la interpretación de los **parámetros estimados** en términos de cocientes de chances, difiere entre los ajustes. Los estimadores de un modelo marginal describen los cocientes de chances en la población, mientras que los correspondientes a los modelos mixtos describen los cocientes de chances en cada *cluster*. Por lo tanto, la elección del enfoque dependerá, en última instancia, de las preguntas de investigación planteadas.

El **diagnóstico** de un modelo para respuestas binarias correlacionadas no resulta sencillo, máxime si se adopta el enfoque marginal que carece de una función de verosimilitud. Bajo estas condiciones, ningún residuo del cual se conozcan sus características distribucionales y que se encuentre implementado resulta útil. Asimismo, si las covariables son binarias, los gráficos usuales no son informativos.

Aunque no existan técnicas formales o informales para diagnosticar el modelo, sí puede evaluarse su **poder predictivo** con distintos métodos. Mediante el cálculo de la tasa de error aparente, se establece el porcentaje de respuestas incorrectamente clasificadas una vez especificado un criterio de clasificación. Para su determinación es conveniente utilizar información proveniente del modelo, en lugar del valor 0.5 que implica que los costos de mala clasificación son iguales en ambos sentidos. Otra alternativa la constituyen las curvas ROC, las cuales relacionan la sensibilidad y la especificidad para distintos puntos de corte: cuanto mayor es el área ubicada por debajo de la curva, mayor es la capacidad predictiva del modelo.

Otro método consiste en aplicar la técnica de **validación cruzada**, siendo la modalidad *leave-one-out* la más adecuada dada la estructura de *clusters* que poseen los datos. La tasa de error calculada por validación cruzada se compara con la tasa de error aparente, a partir de lo cual se concluye acerca del poder predictivo del modelo.

En lugar de realizar inferencia basada en modelos, puede emplearse la **inferencia clásica** o basada en diseño muestral con el fin de estimar, a partir de las encuestas, la proporción de individuos que presentan la característica de interés. Sin embargo, las fórmulas a utilizar para hallar estimaciones de la media y de la varianza requieren disponer de información poblacional, lo cual limita las posibilidades de análisis.

Si se quieren comparar las proporciones estimadas en distintas subpoblaciones, es necesario efectuar una estratificación a posteriori, lo cual reduce el tamaño de la muestra. Como consecuencia, se incrementan la varianza y la amplitud de los intervalos de confianza, tornándose la inferencia menos precisa. Ello desalienta utilizar más de una vía de clasificación, aún en el caso hipotético de disponer de información externa acerca de las covariables. Además, el número de *clusters* efectivamente utilizados depende de que éstos

contengan más de un individuo que posea las características deseadas y que haya variabilidad suficiente para estimar la varianza.

Para **comparar ambos métodos de inferencia** –basada en diseño y basada en modelos–, las covariables ignoradas por la inferencia clásica deben reemplazarse en el predictor lineal de los modelos por sus respectivas proporciones muestrales. Si los *clusters* son de tamaño desigual, las medias estimadas basadas en modelos se consideran más informativas, ya que no se encuentran ponderadas por el tamaño de los *clusters*.

La **precisión de la inferencia** suele ser mayor para el modelo marginal que para los modelos mixtos, debido a que éstos exhiben mayor variabilidad al referirse a *clusters* específicos. En cuanto a la inferencia clásica, la precisión se encuentra estrechamente ligada al tamaño de la muestra en la subpoblación de interés que, como ya se mencionara, se ve reducido en el proceso de condicionar sobre distintas covariables.

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. Entre ellas, pueden citarse: la usual falta de marcos muestrales completos, la carencia de información disponible ajena a la muestra y la dificultad que implica determinar qué fórmulas son adecuadas para estimar las proporciones y sus varianzas ante un diseño muestral complejo.

Luego del estudio efectuado, se propone la siguiente **estrategia de análisis** para la estimación de una proporción a partir de un muestreo por conglomerados:

- Plantear con claridad los objetivos que se persiguen para definir correctamente la población objetivo.
- Plantear las preguntas de investigación antes de decidir si estimar un modelo marginal o un modelo mixto.
- Seleccionar una muestra probabilística si se dispone de marcos de información adecuados. De lo contrario, disponer de una muestra no probabilística en tanto la selección de las unidades muestrales no se encuentre correlacionada con la característica a medir.

- Formular un modelo válido, realizando las pruebas necesarias para determinar la bondad del ajuste, efectuando un análisis diagnóstico y evaluando su poder predictivo.
- Si la correlación entre pares de respuestas es baja, estimar un modelo marginal con estructura de correlación de simetría compuesta. Si la correlación es de moderada a alta, probar distintas estructuras de correlación y compararlas entre sí.

La inferencia basada en modelos, cuyo campo tradicional de aplicación es el de las Ciencias Biológicas, también resulta apropiada a las Ciencias Sociales. 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.

10. FUTURAS INVESTIGACIONES

Si bien las técnicas de diagnóstico adecuadas a los modelos lineales generalizados y a los modelos lineales generalizados mixtos –basados en la función de verosimilitud– se hallan desarrolladas e incorporadas en los paquetes estadísticos que permiten ajustar estos modelos, no sucede lo mismo con los modelos marginales. A pesar de la intensa revisión de trabajos recientes y del contacto con autores de algunos artículos teóricos, los métodos de selección y los estadísticos de bondad del ajuste propuestos para los modelos basados en la función de cuasi-verosimilitud no han sido aplicados, excepto en ejemplos triviales. Queda planteada la relevancia de que investigaciones futuras se dirijan a la implementación de los algoritmos para el diagnóstico de los modelos estimados por el método de ecuaciones de estimación generalizadas, a fin de contar con herramientas formales habida cuenta de la gran flexibilidad y el amplio campo de aplicabilidad que ellos presentan.

Asimismo, aún falta desarrollar técnicas formales e informales de diagnóstico adecuadas para variables de naturaleza binaria. La totalidad de las herramientas gráficas disponibles para el diagnóstico de un modelo suponen que la variable respuesta o que las covariables son continuas, como así también los residuos de *deviance* y de Anscombe, que se emplean en esta fase. Dadas las numerosas aplicaciones en las que las variables respuesta poseen esta distribución, es necesario que se investigue en este sentido.

Por último, los métodos de *bootstrap* adecuados para la estimación de la tasa de error, presentan complicaciones adicionales si el muestreo no es aleatorio irrestricto. En tales casos, es necesario que el remuestreo se realice al interior de cada *cluster* a fin de no ejercer una influencia no deseada sobre la estructura de *clusters* existente en la muestra original.

11. BIBLIOGRAFÍA

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

——— (1996), *An introduction to categorical data analysis.* New York: John Wiley.

Bao, H. [2000], "Evaluation of discovered knowledge" [en línea]. Ch. 7, In: *Knowledge discovery and data mining techniques and practice*" <http://www.netnam.vn/unescocourse/knowlegde/knowlegd.htm> [Consulta: 5 feb. 2004].

Beitler, P. & Landis, J. (1985), "A mixed-effects model for categorical data". *Biometrics*, 41: 991–1000.

Brewer, K. (1999), "Design-based or prediction-based inference? Stratified random vs stratified balanced sampling". *International Statistical Review*, 67 (1): 35–47.

Brier, S. (1980), "Analysis of contingency tables under cluster sampling". *Biometrika*, 67 (3): 591–596.

Cantor, A. (1997), *Extending SAS survival analysis techniques for medical research*. Cary, NC: SAS Institute Inc.

Cochran, W. (1980), *Técnicas de muestreo*. México: CECSA.

Collett, D. (1991), *Modelling binary data.* London: Chapman & Hall.

——— (1994), *Modelling survival data in medical research.* London: Chapman & Hall.

Cornfield, J. (1951), "Modern methods in the sampling of human populations: the determination of sample size". *American Journal of Public Health*, 41: 654–661.

Côté, M. (1991), *By way of advice growth strategies for the market driven world.* Oakville: Mosaic Press.

Cox, D. & Snell, E. (1968), "A general definition of residuals". *Journal of the Royal Statistical Society,* Ser. B, 30 (2): 248–275.

Díaz, M. y Demetrio, C. (1998), *Introducción a los modelos lineales generalizados: Su aplicación a las ciencias biológicas*. Córdoba: Screen.

Diggle, P. ; Heagerty, P. ; Liang, K. & Zeger, S. (2002), *Analysis of longitudinal data.* 2nd ed. New York: Oxford University Press.

Dobson, L. (1983), *An introduction to statistical modelling.* New York: Chapman & Hall.

Dunn, P. & Smyth, G. (1996), “Randomized quantile residuals”. *Journal of Computational and Graphical Statistics*, 5: 236–244.

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

Gennero, A. ; Baltar, F. y Liseras, N. (1999), “Diferencias espaciales en la gestación de ideas empresariales en la Argentina” [en cd-rom]. *Anales del IV Seminario de Red Pymes Mercosur*. Fortaleza, Brasil.

Gilchrist, R. & Green, P. (1996), “The theory of generalized linear models”, in: Francis, B. ; Green, M. & Payne, C. (eds) *The GLIM system: Released 4 manual*, pp. 259–305. Oxford: Clarendon Press.

Gill, J. (2001), *Generalized linear models: A unified approach*. Sage University Papers Series on Quantitative applications in the social sciences. Thousand Oaks, CA: Sage.

Hand, D. (1996), *Construction and assessment of classification rules*. 3rd ed. New York: John Wiley.

Henrekson, M. & Rosenberg, N. (2001), “Designing efficient institutions for science-based entrepreneurship: lessons from the US and Sweden”. *Journal of Technology Transfer*, 26 (3): 207–231.

Hidiroglou, M. & Rao, J. (1987a), “Chi-square tests with categorical data from complex surveys: Part I”. *Journal of Official Statistics*, Statistics Sweden, 3 (2): 117–132.

——— (1987b), “Chi-square tests with categorical data from complex surveys: Part II”. *Journal of Official Statistics*, Statistics Sweden, 3 (2): 133–140.

Hisrich, R. (1988), “Entrepreneurship: Past, present and future”. *Journal of Small Business Management*, 26 (4): 1-4.

Kish, L. (1965), *Survey sampling*. New York: John Wiley.

Kmenta, J. (1977), *Elementos de econometría*. España: Vinces Vives.

Larsen, K. *et al.* (2000), “Interpreting parameters in the logistic regression model with random effects”. *Biometrics*, 56: 909–914.

Levy, P. & Lemeshow, S. (1999), *Sampling of populations: Methods and applications.* 3rd ed. New York: John Wiley.

Liang, K. & Zeger, S. (1986), "Longitudinal data analysis using generalized linear models". *Biometrika*, 73 (1): 13–22.

Lindstrom, M. & Bates, D. (1990), "Nonlinear mixed effects models for repeated measures data". *Biometrics*, 46: 673–687.

Lipsitz, S. *et al.* (1994), "Performance of generalized estimating equations in practical situations". *Biometrics*, 50: 270–278.

Littell, R. *et al.* (1996), *SAS system for mixed models*. Cary, NC: SAS Institute Inc.

Longford, N. (1994), "Logistic regression with random coefficients". *Computational Statistics and Data Analysis*, 17: 1–15.

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

McCulloch, C. & Searle, S. (2001), *Generalized, linear and mixed models.* New York: John Wiley.

Milliken, G. & Johnson, D. (1992), *Analysis of Messy Data - Vol. I: Designed Experiments.* New York: Chapman & May.

Moisen, G. ; Cutler, R. & Edwards, T. (1999), "Generalized linear mixed models for analyzing error in a satellite-based vegetation map of Utah", in: Mowrer, H. & Congalton, R. (eds) *Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing.* Chelsea: Ann Arbor Press.

Neuhaus, J. ; Kalbfleisch, J. & Hauck, W. (1991), "A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data". *International Statistical Review*, 59 (1): 25–35.

Palmgren, J. & Ripatti, S. (2002), "Fitting exponential family mixed models". *Statistical Modelling*, 2: 23–38.

Pan, W. (2001), "Model selection in estimating equations". *Biometrics*, 57: 529–534.

Pendergast, J. *et al.* (1996), "A survey of methods for analyzing clustered binary response data". *International Statistical Review*, 64 (1): 89–118.

Reynolds, P. *et al.* (2002), *Global Entrepreneurship Monitor (GEM) 2001- Expanded Research Report.* London Business School and Babson College.

Rodriguez, G. & Goldman, N. (1995), "An assessment of estimation procedures for multilevel models with binary response". *Journal of the Royal Statistical Society*, Ser. A, 158: 73–89.

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

——— (1999b), "The four types of estimable functions" [en cd-rom], ch. 12, in: *SAS/STAT User Guide.* Cary, NC: SAS Institute Inc.

Schall, R. (1991), "Estimation in generalized linear models with random effects". *Biometrika*, 78: 719–727.

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

Scott, M. & Twomey, D. (1988), "The long-term supply of entrepreneurs: student's career aspirations in relation to entrepreneurship". *Journal of Small Business Management*, 26 (4): 5–13.

Smyth, G. (2003), "Pearson's goodness of fit statistic as a score test statistic", in: Goldstein, D. (ed) *Science and Statistics: A Festschrift for Terry Speed*, IMS Lecture Notes - Monograph Series, 40: 115–126. Institute of Mathematical Statistics, Beachwood, Ohio,

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): 439–455.

Swan, A. (1986), "The use of the deviance to test the goodness of fit of a logistic-linear model to binary data". *Proc GLIM Conference*, London.

Zeger, S. & Liang, K. (1986), "Longitudinal data analysis for discrete and continuous outcomes". *Biometrics*, 42: 121–130.

Zeger, S. ; Liang, K. & Albert, P. (1988), "Models for longitudinal data: a generalized estimating equation approach". *Biometrics*, 44: 1049–1060.

2002, "What are cross-validation and bootstrapping?" [en línea], Part 3. Usenet newsgroup comp.ai.neural-nets. Archivo en: <http://www.faqs.org/faqs/ai-faq/neural-nets/part3/section-12.html> [Consulta: 6 feb. 2004].

12. ANEXO A

12.1. Formulario de encuesta

Universidad Nacional de Mar del Plata

Facultad de Ciencias Económicas y Sociales

ALUMNOS UNIVERSITARIOS: SUS PERCEPCIONES LABORALES

OBJETIVO:
La presente encuesta es parte de un proyecto de investigación que lleva a cabo el Grupo de Análisis Industrial de la Facultad de Ciencias Económicas y Sociales de la Universidad Nacional de Mar del Plata, el cual pretende analizar las percepciones laborales de los alumnos a punto de egresar.
INSTRUCCIONES:

- La encuesta es voluntaria y anónima.
- Lea atentamente todas las opciones antes de contestar. Si debe elegir una sola respuesta, elija aquélla con la que más se identifique usted mismo.
- Ante cualquier duda consulte a la persona a cargo.

A) DATOS PERSONALES:

1. Universidad/Facultad a la que asiste: _____

2. Carrera que cursa actualmente: _____

3. Género:
☐ Hombre ☐ Mujer

4. Edad: _____ años

5. Estado civil:
☐ Soltero ☐ Casado / En pareja ☐ Separado / Viudo

6. ¿Se ha mudado de ciudad para estudiar?
☐ Sí ☐ No

7. Año en inició sus estudios universitarios: _____

8. ¿Alguna vez ha iniciado otra carrera terciaria y/o universitaria?
☐ No
☐ Sí → *8.1.* ¿La ha finalizado? Sí ☐ No ☐

9. Actualmente, ¿se encuentra trabajando *sea con o sin remuneración*?:
☐ No → *9.1.* ¿Alguna vez trabajó *con o sin remuneración*? ☐ Sí ☐ No
☐ Sí → *9.2.*
☐ En relación de dependencia en el sector público
☐ En relación de dependencia en el sector privado en una pyme
☐ En relación de dependencia en el sector privado en una gran empresa
☐ En una empresa perteneciente a familiares
☐ En una empresa propia
☐ Otro: _____

10. ¿Alguna vez realizó una pasantía en una empresa?
☐ No
☐ Sí →
10.1. ¿Qué tipo de empresa era? Pyme ☐ Grande ☐ Pública ☐ Privada ☐
10.2. ¿En qué sector se desempeñaba la empresa? Industria ☐ Comercio ☐ Servicios ☐

11. Actualmente, ¿se encuentra buscando activamente trabajo?: (leyendo los avisos clasificados, concertando entrevistas, presentando el curriculum, etc.)
☐ Sí ☐ No

B) PERSPECTIVAS LABORALES:

12. Al graduarse, a usted **le gustaría** (en el país o en el exterior): *elija una sola respuesta*

- ☐ Continuar estudiando (postgrado, maestría, doctorado)
- ☐ Buscar/continuar trabajando en relación de dependencia en el sector público
- ☐ Buscar/continuar trabajando en relación de dependencia en el sector privado en una pyme
- ☐ Buscar/continuar trabajando en relación de dependencia en el sector privado en una gran empresa
- ☐ Crear su propia empresa (actividad primaria, industria, comercio, servicios)
- ☐ Otro: _____

13. Cree que **ganaría más dinero**: *elija una sola respuesta*

- ☐ Trabajando en relación de dependencia en el sector público
- ☐ Trabajando en relación de dependencia en el sector privado en una pyme
- ☐ Trabajando en relación de dependencia en el sector privado en una gran empresa
- ☐ Trabajando en una empresa propia
- ☐ Otro: _____

14. Cree que su formación universitaria **lo prepara fundamentalmente para**: *elija una sola respuesta*

- ☐ Trabajar en relación de dependencia en el sector público
- ☐ Trabajar en relación de dependencia en el sector privado en una pyme
- ☐ Trabajar en relación de dependencia en el sector privado en una gran empresa
- ☐ Crear su propia empresa
- ☐ Otro: _____

15. Si al graduarse usted no encuentra una oportunidad laboral que se adecue a sus capacidades, ¿cuál piensa que sería su actitud durante en el **corto plazo**?: *elija una sola respuesta*

- ☐ Aceptaría un trabajo para el cual me encuentro sobrecalificado
- ☐ Aceptaría un trabajo no relacionado con mi profesión
- ☐ Buscaría alguna idea u oportunidad de negocios para iniciar una empresa propia
- ☐ Seguiría buscando trabajo que se adecue a mis capacidades
- ☐ Otro: _____

16. ¿Usted percibe que la educación universitaria que ha adquirido le permitirá en el **mediano plazo**:

- ☐ Percibir mayores ingresos trabajando en relación de dependencia que como auto-empleado
- ☐ Percibir mayores ingresos trabajando como auto-empleado que en relación de dependencia

C) PERCEPCION DE LA ACTIVIDAD EMPRESARIAL:

17. ¿Alguna vez asistió a un curso sobre creación de empresas?:

- ☐ No
- ☐ Sí → *17.1.* ¿Fue organizado por su facultad/universidad? Sí ☐ No ☐
  17.2. ¿Satisfizo sus expectativas? Sí ☐ No ☐

18. ¿Sabe si en su facultad/universidad se dictan cursos sobre dicha temática?:

- ☐ No
- ☐ Sí → *18.1.*¿Cuál/es? _____

19. ¿Considera que su carrera le ha dado las herramientas necesarias para crear una empresa propia en el **mediano plazo**?

☐ Sí ☐ Parcialmente ☐ No

Se denomina EMPRENDEDOR a quien lleva a adelante un emprendimiento. Un EMPRENDIMIENTO es el proceso de iniciación de un proyecto de negocios o empresarial y la organización de los recursos necesarios para su puesta en marcha.

20. Usted considera que las personas que crean una empresa: *elija una sola respuesta*

☐ Nacen emprendedores ☐ Se hacen emprendedores ☐ Ambos

21. Su padre se desempeña o se ha desempeñado laboralmente:

- *21.1.* ☐ En relación de dependencia en el sector público
- *21.2.* ☐ En relación de dependencia en el sector privado en una pyme
- *21.3.* ☐ En relación de dependencia en el sector privado en una gran empresa

21.4. ☐ En una empresa propia
21.5. ☐ Otro: _____

22. ¿ Tiene algún otro familiar o amigo cercano que sea empresario?:
☐ No ☐ Sí ⟶ *22.1.* ☐ Otro familiar
22.2. ☐ Un amigo cercano

23. ¿Alguna vez **pensó** en crear su propia empresa?:
☐ No
☐ Sí ⟶ *23.1.* ¿Tiene actualmente un proyecto concreto? Sí ☐ No ☐
23.1.1. ¿El proyecto se vincula con su profesión? Sí ☐ No ☐

24. ¿ Alguna vez **inició** una empresa propia?
☐ No
☐ Sí ⟶ *24.1.* Describa brevemente el emprendimiento:

_____

_____

24.2. ¿Aún continúa en funcionamiento? Sí ☐ No ☐

25. ¿Cómo se valora a usted mismo en los siguientes aspectos, respecto de las personas que habitualmente lo rodean?

	Por encima	**Igual**	**Por debajo**
25.1. Optimismo			
25.2. Capacidad de negociación			
25.3. Creatividad			
25.4. Capacidad de trabajo			
25.5. Capacidad de aprendizaje			
25.6. Capacidad de asumir riesgos			
25.7. Reflexión antes de tomar una decisión			

26. Califique las siguientes afirmaciones <u>*con un círculo*</u>: (1) Completamente de **acuerdo**; (2) Parcialmente de acuerdo; (3) Ni de acuerdo ni en desacuerdo; (4) Parcialmente en desacuerdo; (5) Totalmente en **desacuerdo.**

	+A				**+D**
26.1. Yo puedo influir en mis posibilidades de éxito en lo que hago.	**1**	**2**	**3**	**4**	**5**
26.2. Cuando logro el objetivo propuesto, me fijo nuevas metas más desafiantes.	**1**	**2**	**3**	**4**	**5**
26.3. Me desanimo fácilmente si no estoy seguro de poder cumplir mis metas	**1**	**2**	**3**	**4**	**5**
26.4. Prefiero arriesgarme a perder y no lamentar haber perdido una oportunidad.	**1**	**2**	**3**	**4**	**5**
26.5. Si tengo que tomar una decisión difícil, tiendo a posponerla.	**1**	**2**	**3**	**4**	**5**
26.6. Para mí es fundamental que un trabajo implique constantes desafíos.	**1**	**2**	**3**	**4**	**5**
	+A				**+D**
26.7. Los medios de comunicación local hacen un buen trabajo de cobertura de las noticias sobre empresas.	**1**	**2**	**3**	**4**	**5**
26.8. La gente joven es estimulada a ser independiente y comenzar un nuevo negocio.	**1**	**2**	**3**	**4**	**5**
26.9. Aquéllos que son empresarios exitosos llaman la atención y son admirados	**1**	**2**	**3**	**4**	**5**
26.10. Me gustaría que mis hijos o familiares cercanos fuesen empresarios	**1**	**2**	**3**	**4**	**5**

27. Señale si en su tiempo libre realiza alguna de las siguientes actividades: <u>*puede elegir más de una respuesta*</u>
☐ Reuniones frecuentes con amigos
☐ Práctica de algún deporte
☐ Participación en alguna asociación religiosa, política, ecologista, etc.
☐ Realización de alguna actividad creativa (pintura, música, escritura, etc.)
☐ Hobby ⟶ *27.1.* Cuál?_____

28. **"La idea de iniciar mi propia empresa..."**: <u>*elija una sola opción*</u>
☐ Me da miedo
☐ Me resulta atractiva
☐ No me interesa en absoluto

29. Por favor, diga en qué estrato se ubica el nivel de **ingresos mensuales** de su hogar:

- ☐ Hasta $500
- ☐ $500 a $1.000
- ☐ $1.000 a $3.000
- ☐ Más de $3.000

Gracias por su colaboración!

✓ Le agradeceríamos que nos diese sus datos para poder contactarlo en el futuro:

Nombre y Apellido: _____

Dirección de correo electrónico:_____

Domicilio:_____

Teléfono:_____ _

✓ Si así lo desea, háganos sus comentarios o sugerencias:

12.2. Definición de las variables

En la Tabla A-1 se definen las variables que surgen de la encuesta, identificándolas con el número de pregunta correspondiente.

TABLA A-1: Definición de variables de la encuesta

Número pregunta	Variable
1	Universidad
2	Titulación cursada
3	Género
4	Edad
5	Estado Civil
6	Si es migrante
7	Año inicio estudios universitarios
8	Si inició otra carrera universitaria
8.1	Si finalizó otra carrera universitaria
9	Condición de actividad
9.1	Si alguna vez trabajó
9.2	Lugar de trabajo
10	Realización de pasantías
10.1	Tipo de empresa en la que realizó la pasantía
10.2	Sector económico en el que realizó la pasantía
11	Si actualmente está buscando trabajo
12	Preferencias al graduarse
13	Comparación de ingresos relativos
14	Orientación de la carrera
15	Actitud en el corto plazo frente a una situación de desempleo
16	Comparación de ingresos relativos debidos a la formación universitaria
17	Si realizó algún curso sobre creación de empresas
18	Si en la universidad se dictan cursos sobre creación de empresas
19	Si la formación universitaria le brindó las herramientas necesarias para crear una empresa
20	Posibilidad de influir sobre la vocación emprendedora
21	Experiencia laboral del padre
22	Modelos de rol de familiares o amigos
23	Si alguna vez pensó en crear una empresa
23.1*	Si tiene un proyecto concreto
23.1.1	Si el proyecto se vincula con la carrera
24*	Si alguna vez inició una empresa propia
24.1	Descripción del emprendimiento

Número pregunta	Variable
24.2	Si el emprendimiento continúa en funcionamiento
25.1	Nivel de optimismo
25.2	Capacidad de negociación
25.3	Nivel de creatividad
25.4	Capacidad de trabajo
25.5	Capacidad de aprendizaje
25.6	Capacidad de asumir riesgos
25.7	Reflexión antes de tomar una decisión
26.1	Autoconfianza
26.2	Autoconfianza
26.3	Autoconfianza
26.4	Propensión al riesgo
26.5	Propensión al riesgo
26.6	Propensión al riesgo
26.7	Cultura empresarial
26.8	Cultura empresarial
26.9	Cultura empresarial
26.10	Cultura empresarial
27	Actividades realizadas en el tiempo libre
28	Visualización de la actividad emprendedora
29	Nivel de ingresos mensuales

* Indica las preguntas que se han utilizado para determinar la presencia de vocación emprendedora en los alumnos encuestados.

12.3. Unidades muestrales

En la Tabla A-2 se detallan las universidades y facultades que han sido muestreadas, indicando el tipo de gestión pública o privada de las mismas.

TABLA A-2: Universidades y facultades muestreadas

Universidad	Facultad	Gestión
Universidad de Buenos Aires (UBA) Ciudad Autónoma de Buenos Aires	Cs. Económicas Ingeniería	Pública
Universidad Nacional de Luján (UNLU) Sedes: Luján, Campana y Chivilcoy, Gran Buenos Aires	Cs. Económicas	Pública
Universidad Nacional de La Plata (UNLP) La Plata, Pcia. de Buenos Aires	Cs. Económicas Ingeniería	Pública
Universidad Nacional de Mar del Plata (UNMDP) Mar del Plata, Pcia. de Buenos Aires	Cs. Económicas Ingeniería	Pública
Universidad Tecnológica Nacional (UTN) Sede: Avellaneda, Gran Buenos Aires	Ingeniería	Pública
Universidad de Belgrano (UB) Ciudad Autónoma de Buenos Aires	Cs. Económicas Ingeniería	Privada
Universidad de Morón (UNIMORON) Morón, Gran Buenos Aires	Cs. Económicas Ingeniería	Privada
Universidad Centro de Altos Estudios en Ciencias Exactas (CAECE) Sede: Mar del Plata, Pcia. de Buenos Aires	Cs. Económicas	Privada
Universidad Fraternidad de Santo Tomás de Aquino (FASTA) Sede: Mar del Plata, Pcia. de Buenos Aires	Cs. Económicas	Privada

En la Tabla A-3 se incluye el listado de titulaciones muestreadas por facultad, así como el número de encuestas procesadas correspondientes a cada curso, luego de descontar

aquéllas con datos faltantes. Asimismo, se define el criterio con el que han sido conformados los *clusters* y *subclusters*.

TABLA A-3: Titulaciones muestreadas, definición de *clusters* y *subclusters* y número de encuestas procesadas

Cluster	*Subcluster*	Titulación	Encuestas procesadas
1	1	Lic. en Administración	92
	2	Lic. en Economía	45
9	13	Ing. en Sistemas Ing. en Informática	3
	14	Ing. Industrial Ing. Mecánica	45
2	3	Lic. en Administración	46
10	15	Ing. Civil Ing. en Construcciones Ing. Mecánica	18
	16	Ing. Industrial Ing. Química	13
3	4	Lic. en Administración	10
11	17	Ing. en Sistemas de Información Ing. en Informática	13
	18	Ing. Electrónica Ing. Electromecánica Ing. Civil Ing. Industrial	17
4	5	Lic. en Administración	15
	6	Lic. en Economía	6
12	19	Ing. Civil Ing. en Agrimensura	15
5	7	Lic. en Administración	41
	8	Lic. en Economía	27
13	20	Ing. Aeronáutica	9
	21	Ing. Civil Ing. en Vías de Comunicación Ing. en Construcciones Ing. Hidráulica	40
	22	Ing. Electricista Ing. Electrónica	26
	23	Ing. Industrial Ing. Mecánica	45
6	9	Lic. en Administración	87
	10	Lic. en Economía	10
14	24	Ing. Química Ing. en Materiales Ing. en Alimentos Ing. Industrial Ing. Mecánica	42
	25	Ing. Electrónica Ing. Electromecánica Ing. Eléctrica	35
7	11	Lic. en Administración de Negocios	17
8	12	Lic. en Administración	6
		Total	**723**

Nota: con fines de confidencialidad, el orden de las facultades difiere del indicado en la Tabla A-2.

13. ANEXO B

Una alternativa al uso de los estimadores para proporciones, es optar por los estimadores de razón *(ratio)*. Las fórmulas para la varianza resultan entonces más simples, debido a que se alimentan con los totales de los conglomerados de primera etapa, sin que sea necesario descomponer la varianza entre y dentro de las unidades primarias[36]. Sean:

$i = 1,...,k$ universidades
$j = 1,...,m_i$ alumnos .

Si la cantidad que va a ser estimada a partir de una muestra aleatoria es la razón de dos variables, ambas variando de unidad a unidad, el parámetro poblacional es:

$$R = \frac{\sum_{i=1}^{K}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{M_i} x_{ij}}, \qquad \textbf{[B-1]}$$

y el estimador está dado por:

$$r = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m_i} y_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{m_i} x_{ij}}. \qquad \textbf{[B-2]}$$

La distribución derivada del muestreo en muestras pequeñas es asimétrica y algo sesgada, pero en muestras grandes tiende a la normalidad y el sesgo se torna despreciable (Cochran, 1980). Si la estimación de razón se obtiene a partir de un muestreo por conglomerados en dos etapas, la varianza estimada es (Levy & Lemeshow, 1999):

$$Var(r) = r^2\left(\frac{N-n}{Nk}\right)\left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{m_i}\left(y_{ij}-\bar{y}\right)^2}{(k-1)\bar{y}^2} + \frac{\sum_{i=1}^{k}\sum_{j=1}^{m_i}\left(x_{ij}-\bar{x}\right)^2}{(k-1)\bar{x}^2} - 2\frac{\sum_{i=1}^{k}\sum_{j=1}^{m_i}\left(y_{ij}-\bar{y}\right)\left(x_{ij}-\bar{x}\right)}{(k-1)\bar{y}\bar{x}}\right), \qquad \textbf{[B-3]}$$

[36] No obstante, si existe gran heterogeneidad en las probabilidades de selección –explicada por *clusters* de tamaños muy desiguales–, resulta conveniente reponderar los totales de las unidades de primera etapa con sus respectivas inversas.

siendo:

- *N* el total de individuos en la población.
- *n* el total de individuos en la muestra.
- *f* la fracción de muestreo dentro de la unidad primaria.
- $\bar{y}$ el promedio de individuos que presentan el atributo.
- $\bar{x}$ el tamaño medio del *cluster*.

El primer término puede rescribirse como $\left(\frac{1-f}{k}\right)$, lo que implica que la fracción de muestreo debe ser aproximadamente constante. El uso de este estimador para el cálculo de la proporción de alumnos con vocación emprendedora ha sido probado, obteniéndose:

$$\hat{\mu}_{..} = 0.3610$$
$$Var\left(\hat{\mu}_{..}\right) = 0.0009 \quad ,$$
$$IC\left(\hat{\mu}_{..}\right) = (0.30, 0.42)$$

pero, el supuesto acerca de la fracción de muestreo constante invalida su uso.

En la aplicación realizada, no es posible mantener la proporcionalidad entre el tamaño de la muestra por *cluster* y el tamaño del *cluster*, debido a que éste último es altamente variable. Dentro de la población objetivo existen facultades con más de 2000 alumnos cursando el último año de las carreras de interés (Universidad de Buenos Aires), mientras que otras instituciones poseen menos de 10 estudiantes (Universidad Nacional de Gral. Sarmiento)[37]. Por lo tanto, el diseño muestral no puede garantizar una fracción de muestreo constante.

[37] Ambos ejemplos corresponden a facultades de ciencias económicas de gestión pública.

14. ANEXO C

En este Anexo se presentan algunos avances teóricos que podrían contribuir a la selección y al diagnóstico de modelos para respuestas binarias correlacionadas.

14.1. Método de selección para modelos marginales

Pan (2001) propone que el modelo a seleccionar sea aquél que minimiza el sesgo predictivo esperado *(Expected Predictive Bias - EPB)*. Puesto que en la práctica se dispone de una única muestra para determinar tal valor, pueden obtenerse nuevas muestras mediante los métodos de *bootstrap* o de validación cruzada *(Cross Validation - CV)*[38]. Sin embargo, estimar el *EPB* resulta complejo bajo la primera alternativa cuando existe una estructura de *clusters* y el estimador por validación cruzada no suele ser eficiente.

Dado que el uso de *bootstrap* para obtener estimadores reduce el error estándar, una opción propuesta por el citado autor es utilizar un estimador por validación cruzada suavizado por *bootstrap (BCV)*. Con *B* muestras *bootstrap,* el estimador del *EPB* es aproximado por la siguiente expresión:

$$\widehat{EPB}_{BCV} = \frac{\sum_{b=1}^{B} \left| S\left(X^{*b-} \mid \hat{\beta}\left(X^{*b}\right)\right)\right|}{B}, \qquad \text{[C-1]}$$

donde:

- *S* representa un sistema de ecuaciones de estimación *S(X/β)*.
- *X* representa un vector fila compuesto por el valor observado de la variable respuesta y de las covariables, i.e., una observación completa.
- X^{*b} representa las observaciones incluidas en la muestra *bootstrap*.

[38] Los métodos de *bootstrap* consisten, básicamente, en extraer muestras de tamaño *n* con reemplazo a partir de la muestra original. Los métodos de validación cruzada consisten en extraer subconjuntos de datos a fin de evaluar la regla de clasificación generada con los datos restantes; el procedimiento se repite para distintos subconjuntos y se promedian los resultados (Hand, 1996).

- X^{*b-} representa las observaciones no incluidas en la muestra *bootstrap* y que, por consiguiente, serán utilizadas para la validación.

Para juzgar el modelo bajo análisis, es necesario combinar la evidencia de cada componente del estimador. Idealmente, si cada componente de $\widehat{EPB}_{BCV}$ para un modelo dado es mínimo comparado con los componentes de otro modelo, la elección resulta sencilla. Sin embargo, en la práctica es posible que tal modelo no exista. Asimismo, los distintos componentes de $\widehat{EPB}_{BCV}$ pueden no encontrarse en la misma escala, por lo que resulta conveniente estandarizar las covariables previo al ajuste.

En general, puede emplearse una suma ponderada de los componentes de $\widehat{EPB}_{BCV}$ como estadístico resumen. Los pesos pueden ser inversamente proporcionales a las varianzas de cada componente, estimadas directamente a partir de las muestras *bootstrap*. Aunque la elección de los ponderadores puede afectar la eficiencia del método propuesto, asintóticamente, cada componente de $\widehat{EPB}_{BCV}$ tiende a cero bajo el modelo correcto.

El estadístico a utilizar como criterio de selección del modelo podría construirse de dos formas. La primera consiste en utilizar el componente de $\widehat{EPB}_{BCV}$ correspondiente a la covariable de mayor interés o a la covariable incluida en todos los modelos a comparar, optándose por aquel modelo en el cual exhiba el menor valor (BCV_1). La segunda forma es contemplar el modelo con más términos y comparar los promedios ponderados de todos los componentes para cada uno de los modelos (BCV_a).

Un estudio de simulación referido por Pan (2001), señala a BCV_a como la medida que exhibe una mejor *performance*, lo cual implica que utilizar sólo un subconjunto de los componentes de $\widehat{EPB}_{BCV}$ conlleva una pérdida de información y resulta menos efectivo, aunque dicha pérdida no sea sustancial. La principal ventaja del criterio de selección presentado es que puede utilizarse para comparar modelos no anidados.

Este método aún no ha sido suficientemente evaluado cuando se dispone de observaciones agrupadas. En tal caso, la obtención de muestras *bootstrap* debiera respetar la estructura de conglomerados.

14.2. Residuos cuantiles aleatorizados

Un nuevo tipo de residuo propuesto por Dunn & Smyth (1996) son los residuos de cuantiles aleatorizados *(randomized quantile residuals)*. Éstos son continuos aún si la variable respuesta es de naturaleza discreta, ya que se computan buscando el desvío normal estándar equivalente para cada una de las observaciones. Si bien podría optarse por cualquier otra distribución, la asimetría resulta una complicación innecesaria y las distribuciones restringidas introducen patrones espurios que dificultan la interpretación gráfica.

La función de distribución acumulada se expresa como $F(y;\ \mu,\ \phi)$. Si ésta es continua, entonces se halla uniformemente distribuida en el intervalo [0, 1] y los residuos se definen como:

$$r_{q,i} = \Phi^{-1}\left\{F\left(y_i;\hat{\mu}_i,\hat{\phi}_i\right)\right\}, \qquad \text{[C-2]}$$

siendo Φ^{-1} la función de distribución acumulada normal estándar. Excepto por la variabilidad muestral, estos residuos se distribuyen como una normal estándar, lo que implica que su distribución converge a $N(0,\ 1)$ si los parámetros se estiman en forma consistente.

Si $F(y;\ \mu,\ \phi)$ no es continua, se requiere una definición más general para los residuos. Dados a_i y b_i:

$$a_i = \lim_{y\uparrow y_i} F\left(y;\hat{\mu}_i,\hat{\phi}\right), \qquad \text{[C-3]}$$

$$b_i = F\left(y_i;\hat{\mu}_i,\hat{\phi}\right), \qquad \text{[C-4]}$$

los residuos de cuantiles aleatorizados para y_i –cuya distribución es nuevamente normal estándar, excepto por la variabilidad muestral– se definen como:

$$r_{q,i} = \Phi^{-1}\left(u_i\right), \qquad \text{[C-5]}$$

$$u_i \sim U(a_i,b_i]. \qquad \text{[C-6]}$$

Si bien debido a la aleatorización los residuos van a variar de una realización a otra, para un mismo cuerpo de datos y modelo ajustado, los autores proponen que se grafiquen cuatro realizaciones. Cualquier patrón observado que no sea consistente en los distintos gráficos debe ignorarse.

Aunque Dunn & Smyth (1996) asumen en su trabajo que las observaciones son independientes, comentan que el método puede ser extendido a situaciones en las que existe dependencia si se expresa la función de log-verosimilitud multivariada como una suma de log-verosimilitudes condicionales. Por consiguiente, el método puede ser adaptado a los enfoques basados en la función de verosimilitud –i.e., modelos mixtos–, pero no así al enfoque marginal. No obstante, aún no se ha hecho el desarrollo necesario para su implementación.

15. ANEXO D

15.1. Comandos SAS

Tres procedimientos distintos de SAS son necesarios para estimar los modelos formulados. A continuación se detallan los comandos utilizados en cada caso y se presentan las rutinas que se han corrido para obtener las estimaciones.

Modelo marginal

PROC GENMOD: es el procedimiento que utiliza SAS para estimar un modelo lineal generalizado marginal.

CLASS: especifica las covariables categóricas utilizadas en el análisis.

MODEL: especifica la variable respuesta y las covariables incluidas en el predictor lineal.

DIST: especifica la distribución de probabilidad de la variable respuesta.

LINK: especifica la función de enlace utilizada.

REPEATED: identifica la estructura de dependencia para las respuestas multivariadas.

SUBJECT: es la unidad de datos que se considera dependiente, la cual debe ser especificada en el comando CLASS.

TYPE: especifica la estructura propuesta para la matriz de correlación de trabajo.

- IND: independencia.
- EXCH: simetría compuesta.

LOGOR: especifica la estructura propuesta para los logaritmos de los cocientes de chances entre observaciones del mismo *cluster*.

- EXCH: simetría compuesta.
- NEST1: logaritmos de cocientes de chances anidados a un nivel.
- LOGORVAR: logaritmos de cocientes de chances por *cluster*.

Modelo mixto con verosimilitud completa

PROC NLMIXED: es el procedimiento que utiliza SAS para estimar un modelo de efectos aleatorios basado en la función de verosimilitud completa.

PARMS: lista los parámetros y especifica sus valores iniciales.

MODEL: especifica la distribución condicional de probabilidad de la variable respuesta.

RANDOM: define al efecto aleatorio y especifica su distribución de probabilidad.

Modelo mixto con verosimilitud condicional

PROC PHREG: es el procedimiento que utiliza SAS para estimar un modelo de efectos mixtos basado en la función de verosimilitud condicional, apropiado para estimar el modelo de Cox de riesgo proporcional *(proportional hazard)*.

MODEL: identifica la variable a ser usada como “tiempo de falla” y las covariables.

TIES: especifica cómo tratar las ligas en los tiempos de falla.

- DISCRETE: indica que se utilice el modelo logístico discreto.

STRATA: identifica la variable que agrupa a las observaciones o *cluster*.

15.2. Rutinas SAS

Ajuste del modelo

```
proc genmod data=base descending;
class cluster genero ocupado actitud vision riesgo creativ;
model ve = genero ocupado actitud vision riesgo creativ
         / dist=bin link=logit type3;
repeated subject=cluster / type=exch; run;

proc nlmixed data=base;
parms beta0=-3.9  beta1=0.8   beta2=1.0   beta3=1.1
      beta4=1.6   beta5=0.8   beta6=0.5   sigma2=0.05;
pred= beta0+beta1*genero+beta2*ocupado+beta3*actitud+beta4*vision+
      beta5*riesgo+beta6*creativ+u;
prob= exp(pred)/(1+exp(pred));
model ve ~ binary(prob);
random u ~ normal(0,sigma2) subject=cluster;
predict exp(pred)/(1+exp(pred)) out=predicciones;
```

```
predict u out=efectoscluster; run;

proc phreg;
model time*ve(0)= genero ocupado actitud vision riesgo creativ
      / ties=discrete rl;
strata cluster; run;
```

Contrastes

```
proc genmod data=a;
(...)
contrast 'genero=0 ocupado=0' genero -1 1, ocupado -1 1;
contrast 'actitud=0 vision=0' actitud -1 1, vision -1 1;
contrast 'riesgo=0 creativ=0' riesgo -1 1, creativ -1 1;
contrast 'todas 0' genero 1 -1, ocupado 1 -1, actitud 1 -1, vision 1 -1,
riesgo 1 -1, creativ 1 -1; run;

proc nlmixed data=a;
(...)
contrast 'beta1=0, beta2=0' beta1, beta2;
contrast 'beta1=beta2' beta1-beta2; run;
```

Tasa de error aparente

```
(...)
output out=predicciones predicted=predichos;
data predicciones;
set predicciones;
z=0.4;
prob=(predichos>=z);
proc freq;
table ve*prob / nopercent nocol; run;
```

Macro curvas ROC

```
%macro roc(numpuntos);

%do i=0 %to &numpuntos;

data predic;
set predicciones;
z=&i/&numpuntos;
prob=(predichos>=z);
keep ve prob z;
%if &i=1 %then %do; data todo;
                    set predic;
                     %end;
            %else %do; data todo;
                       set todo predic;
                     %end;
%end;
%mend;

roc(40);
```

Macro validación cruzada "leave-one-out"

```
%macro crossv(nobs,base);
ods listing close;

data cval;
 predichos=.;

data betas;
 estimate=.;
 stderr=.;

%do k=1 %to &nobs;

data workds;
 set &base;
 if _n_=&k then ve=.;

proc genmod data=workds descending;
class cluster subcluster genero ocupado actitud vision riesgo creativ;
model ve= genero ocupado actitud vision riesgo creativ
        /dist=bin link=logit type3;
repeated subject=cluster / type=exch;
output out=pred predicted=predichos upper=a2 lower=a3;
ods output parameterestimates=parms;
run;

data parms;
 set parms;
 (...)
 keep estimate stderr;
proc datasets;
 append base=betas data=parms force;

data pred;
 set pred;
 keep predichos;
proc datasets;
 append base=cval data=pred force;
run;

%end;
ods listing;
%mend;

%crossv(723,base);
```